LIDS-P-2278

On the Large Deviations Behaviour of Acyclic Networks of G/G/1 Queues [1]

Dimitris Bertsimas Ioannis Ch. Paschalidis
John N. Tsitsiklis

Laboratory for Information and Decision Systems
and
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139

December 1994

[1]Research supported by a Presidential Young Investigator award DDM-9158118 with matching funds from Draper Laboratory, and by the ARO under grant DAAL-03-92-G-0115.

Abstract

We consider a single class, acyclic network of G/G/1 queues. We impose some mild assumptions on the service and external arrival processes and we characterize the large deviations behaviour of all the processes resulting from various operations in the network. For the network model that we are considering, these operations are passing-through-a-single-server-queue (the process resulting from this operation being the departure process), superposition of independent processes, and Bernoulli splitting of a process to a number of processes. We also characterize the large deviations behaviour of the waiting time and the queue length observed by a typical customer in a single server queue. We prove that the assumptions imposed on the external arrival processes are preserved by these operations, and we show how to inductively apply these results to obtain the large deviations behaviour of the waiting time and the queue length in all the queues of the network. Our results indicate how these large deviations occur, by concretely characterizing the most likely path that leads to them.

Keywords: Communication Networks, Large Deviations, Queueing Networks.

1 Introduction

Consider a single class, acyclic network of G/G/1 queues. Customers arrive to the network in a number of independent streams and are treated uniformly by the network. Different streams may share a queue and the first-come first-serve (FCFS) policy is implemented. A customer departing a queue i, is routed to queue j with probability p_{ij} or leaves the network with probability p_{i0}. The routing decisions are assumed to be independent of everything else in the network. The aim of this paper is to derive large deviations results for the waiting time and the queue length observed by an arbitrary customer at different queues of the network.

The main application area that motivates the study of such systems is the design and the operation of high speed, packet-switched communication networks. These networks will accommodate various types of traffic, namely, digitized voice, encoded video, and data. The interesting problem arising is how to estimate and prevent congestion, which may cause long delays and packet losses. It is desirable to operate the network in a regime where packet loss probabilities are very small, e.g., in the order of 10^{-9}. Moreover, large delays should also have a correspondingly small probability. Thus, the need of understanding the large deviations behaviour of such a network arises. In this paper, we consider single class networks, which from the application point of view means that we are dealing only with one type of traffic in the network. For this reason, the FCFS assumption can be made without loss of generality.

The problem of estimating tail probabilities of rare events in a single queue has received extensive attention in the literature and has been approached by two main methodologies. The first one is to use large deviations theory, as we do in this paper. This approach is used in [dVW92] to estimate the tail probability of the queue length in a G/G/1 queue. In that paper, a discrete time model was used in contrast to the continuous time model that we use in this paper. Similar results are obtained in [CW93]. The second approach is to use spectral decomposition techniques. This second approach is used in [EM93] to estimate the tail probability of the queue length in a queue with a deterministic server and Markov modulated arrival process. Results for the single queue case were first obtained in [Hui88], [Kel91], [GH91] and later in [KWC93]. In all of these papers, the large deviations results obtained are used to derive appropriate admission control schemes for networks.

The extension of these ideas to networks appears to be a rather challenging problem. Researchers have been able to obtain some bounds on the tail probabilities for delays and queue lengths in various networks models (see [Cha94, Cru91a, Cru91b, YS93]), but it is not clear whether these bounds are tight. Recently, large deviations results for two queues in tandem, with renewal arrivals and exponential servers, were reported in [GA94]. In [dVCW93], a very interesting approach is used to obtain results for networks with deterministic servers. The departure process from a single G/D/1 queue is characterized in the large deviations regime, using a discrete time model, in an attempt to treat the whole network inductively.

The main focus of [dVCW93] is to apply the large deviations results obtained to resource management for networks. It is important to point out that the departure process is a very difficult process to obtain exact results for (see for example [BN90]). However, we should note that it is not very clear to us how the large deviations result for the departure process in [dVCW93] can be applied inductively. The crux of the matter is that [dVCW93] uses a technical result from [DZ93a] in order to obtain the large deviations behaviour of the departure process. The latter result holds under certain technical assumptions on the arrival process. Since the departure process from a queue is the arrival process in an other downstream queue in the network, one would need at this point to verify that the same technical assumptions hold for the departure process. This is not done in [dVCW93] and appears to be rather difficult.

In the present paper, we consider a continuous time model and we extend the work in [dVCW93] to a network of G/G/1 queues. Our results are self-contained in the sense that we do not need the technical results of [DZ93a]. Instead, we impose certain assumptions on the external arrival processes and we characterize the large deviations behaviour of all the processes resulting from various operations in the network. For the network model that we are considering, these operations are passing-through-a-queue (the process resulting from this operation being the departure process), superposition of independent processes, and Bernoulli splitting of a process to a number of processes. For a single queue, we characterize the large deviations behaviour of the waiting time incurred by a typical customer and, by using ideas from distributional laws (see [BN91, BM92]), the large deviations behaviour of the queue length observed by a typical customer. We prove that the assumptions imposed on the external arrival processes are preserved by the operations mentioned above, and we are able to apply these results inductively to obtain large deviations results for all the interesting processes in the network. Moreover, our approach provides particular insight on how these large deviations occur, by concretely characterizing the most likely path that leads to them. Characterizations of most likely paths were obtained for the single queue case in [Asm82] and [Ana88].

It is interesting to note, that in order to obtain the large deviations behaviour of the superposition operation we prove a general result that connects the *stationary distribution* (i.e., as it is seen at a random time) and the *Palm distribution* (i.e., as it is seen by a typical customer) of a point process in the large deviations regime. This result could be of independent interest.

Regarding the structure of the paper, we start in Section 2 by reviewing some results from the theory of large deviations that we use in the sequel. In Section 3 we present the network model that we are considering and establish our notation. In Section 4 we treat the single queue case. This section is comprised by two subsections. In Subsection 4.1 we review the existing result for the large deviations behaviour of the waiting time and we completely characterize the most likely path along which the waiting time takes large values.

In Subsection 4.2, using an idea from distributional laws we obtain the tail probability of the queue length. In Section 5 we derive the large deviations behaviour of the departure process from a G/G/1 queue. Particular attention is given to the way that such a deviation occurs. In Subsection 5.1, some special cases are studied. Namely, we apply the result for the departure process of a G/G/1 queue to a G/D/1 queue and an M/M/1 queue. For the latter case, Burke's Theorem is verified in the large deviations regime. In Sections 6 and 7 we study the large deviations behaviour of the processes resulting from the following operations: superposition of independent processes, and Bernoulli splitting of a process to a number of processes, respectively. In Subsection 6.1 we prove a result that connects the Palm and the stationary distribution of a point process in the large deviations regime. This result is used in the rest of Section 6 to derive the large deviations behaviour of the superposition process. In Section 8, we treat, as an example, a network consisting of two queues in tandem. We characterize the way that the waiting time in the second queue reaches large values and we include some numerical results. Finally, in Section 9 we provide some concluding remarks and discuss some open problems.

2 Preliminaries

In this section we review some basic results on Large Deviations Theory that will be used in the sequel.

We first state the Gärtner-Ellis Theorem (see [Buc90] and [DZ93b]) which establishes a *Large Deviations Principle (LDP)* for random variables. It is a generalization of Cramer's theorem which applies to independent and identically distributed (iid) random variables.

Consider a sequence $\{S_1, S_2, \dots\}$ of random variables, with values in $\mathbb{R}$ and define

$$\Lambda_n(\theta) \triangleq \frac{1}{n} \log \mathbf{E}[e^{\theta S_n}]. \tag{1}$$

For the applications that we have in mind, S_n is a partial sum process. Namely, $S_n = \sum_{i=1}^{n} X_i$, where X_i, $i \geq 1$, are identically distributed, possibly dependent random variables.

Assumption A

1. *The limit*

$$\Lambda(\theta) \triangleq \lim_{n\to\infty} \Lambda_n(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{\theta S_n}] \tag{2}$$

   exists for all θ, where $\pm\infty$ are allowed both as elements of the sequence $\Lambda_n(\theta)$ and as limit points.

2. *The origin is in the interior of the domain $D_\Lambda \triangleq \{\theta \mid \Lambda(\theta) < \infty\}$ of $\Lambda(\theta)$.*

3. *$\Lambda(\theta)$ is differentiable in the interior of D_Λ and the derivative tends to infinity as θ approaches the boundary of D_Λ.*

Theorem 2.1 ***(Gärtner-Ellis)*** *Under Assumption A, the following inequalities hold*

Upper Bound: *For every closed set F*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left[\frac{S_n}{n} \in F\right] \leq -\inf_{a\in F} \Lambda^*(a). \tag{3}$$

Lower Bound: *For every open set G*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left[\frac{S_n}{n} \in G\right] \geq -\inf_{a\in G} \Lambda^*(a), \tag{4}$$

where

$$\Lambda^*(a) \triangleq \sup_{\theta}(\theta a - \Lambda(\theta)). \tag{5}$$

We say that $\{S_n\}$ satisfies a LDP with *good rate function* $\Lambda^*(\cdot)$. The term "good" refers to the fact that the level sets $\{a \mid \Lambda^*(a) < k\}$ are compact for all $k < \infty$, which is a consequence of Assumption A (see [DZ93b] for a proof).

It is important to note that $\Lambda(\cdot)$ and $\Lambda^*(\cdot)$ are convex duals (Legendre transforms of each other). Namely, along with (5), it also holds

$$\Lambda(\theta) = \sup_{a}(\theta a - \Lambda^*(a)). \tag{6}$$

The Gärtner-Ellis Theorem intuitively asserts that for large enough n and for small $\epsilon > 0$,

$$\mathbf{P}[S_n \in (na - \epsilon, na + \epsilon)] \sim e^{-n\Lambda^*(a)}.$$

However, in this paper, we are mostly estimating tail probabilities of the form $\mathbf{P}[S_n \leq na]$ or $\mathbf{P}[S_n \geq na]$. We therefore define large deviations rate functions associated with such tail probabilities.

Consider the case where $S_n = \sum_{i=1}^n X_i$, the random variables X_i, $i \geq 1$, being identically distributed, and let $m = \mathbf{E}[X_1]$. It is easily shown (see [DZ93b]) that $\Lambda^*(m) = 0$. Let us now define

$$\Lambda^{*+}(a) \triangleq \begin{cases} \Lambda^*(a) & \text{if } a > m \\ 0 & \text{if } a \leq m \end{cases} \tag{7}$$

and

$$\Lambda^{*-}(a) \triangleq \begin{cases} \Lambda^*(a) & \text{if } a < m \\ 0 & \text{if } a \geq m. \end{cases} \tag{8}$$

The convex duals of these functions are

$$\Lambda^+(\theta) \triangleq \begin{cases} \Lambda(\theta) & \text{if } \theta \geq 0 \\ +\infty & \text{if } \theta < 0 \end{cases} \tag{9}$$

and

$$\Lambda^-(\theta) \triangleq \begin{cases} \Lambda(\theta) & \text{if } \theta \leq 0 \\ +\infty & \text{if } \theta > 0 \end{cases} \tag{10}$$

respectively.

Using the Gärtner-Ellis Theorem we can now state

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_n \leq na] = -\sup_{\theta}(\theta a - \Lambda^-(\theta)) = -\Lambda^{*-}(a) \tag{11}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_n \geq na] = -\sup_{\theta}(\theta a - \Lambda^+(\theta)) = -\Lambda^{*+}(a). \tag{12}$$

3 The Network Model

In this section, we formally define the network model of which we will derive the large deviations behaviour. Moreover, we establish the notation that we will be using and state a set of assumptions on the arrival and service processes.

Consider a *directed acyclic graph (dag)* with J nodes. For reasons that will become soon apparent, we assume that any two directed paths do not meet in more than one nodes. Each node of the graph is equipped with an infinite buffer and a single server. Customers enter the network in a number of independent streams $A^1, A^2, \ldots, A^J$. In particular, A^i is the stream of customers that enter the network at node i. Customers are treated uniformly by the network, i.e., the network is assumed to be *single class*. Let $\mathbb{Z}$ denote the set of integers. By A_i^j, $i \in \mathbb{Z}$, we denote the interarrival time of the ith customer in the jth stream (the interval between the arrival epochs of the $(i-1)$st and the ith customer). By B_i^j, $i \in \mathbb{Z}$, we denote the service time of the ith customer in the jth node. We assume that for each arriving stream j the process $\{A_i^j,\ i \in \mathbb{Z}\}$, is stationary, and A_i^j, $i \in \mathbb{Z}$, are possibly dependent random variables. Moreover, for each node j, the service times B_i^j, $i \in \mathbb{Z}$, are iid random variables. We also assume that interarrival and service times at a specific node are mutually independent and that service times at different nodes are independent.

Independent streams may share a queue and the FCFS policy is implemented. A customer departing node i, connected with nodes $j_1, j_2, \ldots$, is routed to these nodes with probabilities $p_{ij_1}, p_{ij_2}, \ldots$, respectively, or leaves the network with some probability p_{i0}. The routing decisions are assumed to be independent of everything else in the network.

Figure 1 depicts an example of the class of networks considered. Such a network is intended

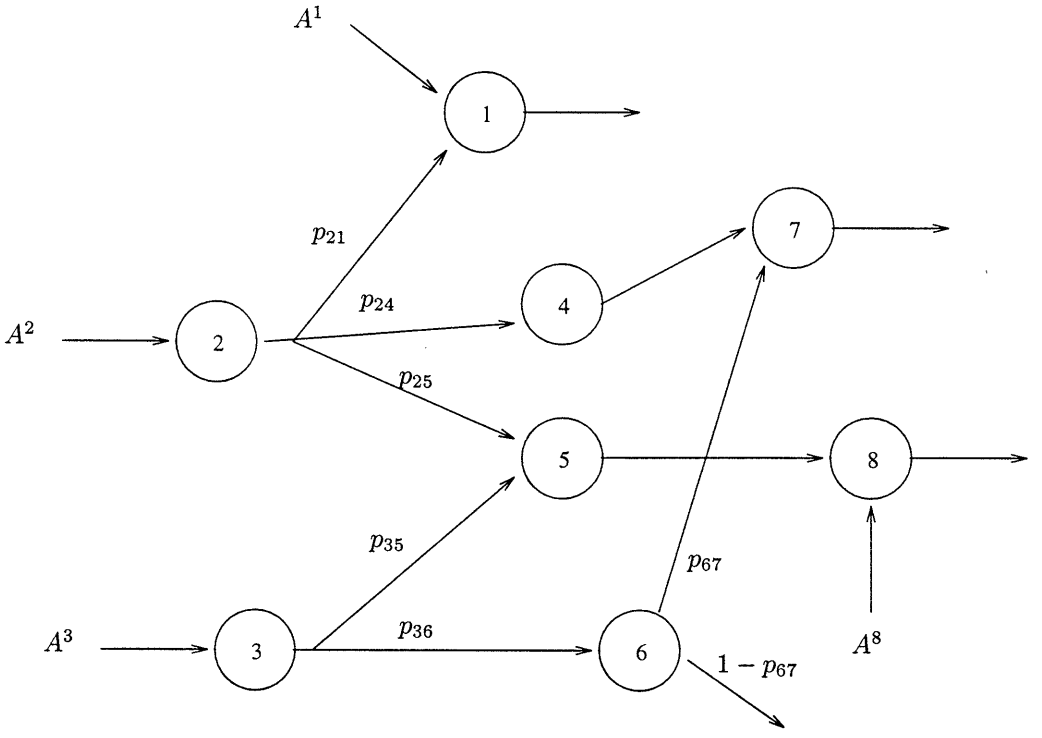

Figure 1: A network example.

to model packet-switched communication networks.

We denote by $W^1, W^2, \ldots, W^J$ and $L^1, L^2, \ldots, L^J$ the steady-state waiting times and queue lengths, incurred by a typical customer at nodes $1, 2, \ldots, J$ of the network, respectively. For each node j, W_n^j (resp. L_n^j) denotes the waiting time incurred (resp. queue length observed) by the nth customer. We assume that the process $\{(W_n^j, L_n^j);\ n \in \mathbb{Z}, j = 1, \ldots, J\}$ is stationary.

In this paper, we derive large deviations results for the steady-state waiting times $W^1, W^2, \ldots, W^J$, and the corresponding queue lengths $L^1, L^2, \ldots, L^J$, incurred at nodes $1, 2, \ldots, J$ of the network, respectively (as these random variables are seen by a typical customer). Our strategy is first to obtain large deviations results for the steady-state waiting time and the corresponding queue length in a single G/G/1 queue. Then it suffices to derive a LDP for the partial sum of the aggregate arrival process in each queue of the network and apply the result for the single queue case. It is important to note that by the definition of the network all the streams sharing the same queue are independent. Therefore, from the model description, it is apparent that it suffices to obtain LDP's for the processes resulting from the following operations

1. Passing-through-a-queue (the process resulting from this operation being the depar-

ture process).

2. Superposition of independent streams.

3. Bernoulli splitting of a stream to a number of streams.

Let $\{A_i,\ i \in \mathbb{Z}\}$ be an arbitrary external arrival process and $\{B_i,\ i \in \mathbb{Z}\}$ be an arbitrary service process. We will be using the notation $S_{i,j}^X \stackrel{\Delta}{=} \sum_{k=i}^{j} X_k;\ i \leq j$ for the partial sums of the random sequence $\{X_i;\ i \in \mathbb{Z}\}$.

Assumption B

1. *The sequence of partial sums* $\{S_{1,n}^A;\ n \geq 1\}$ *satisfies the following LDP*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_{1,n}^A \leq na] = -\Lambda_A^{*-}(a), \tag{13}$$

where

$$\Lambda_A^-(\theta) \stackrel{\Delta}{=} \begin{cases} \lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{\theta S_{1,n}^A}] & \text{if } \theta \leq 0 \\ +\infty & \text{if } \theta > 0 \end{cases} \tag{14}$$

and

$$\Lambda_A^{*-}(a) \stackrel{\Delta}{=} \sup_\theta (\theta a - \Lambda_A^-(\theta)). \tag{15}$$

2. *The sequence of partial sums* $\{S_{1,n}^B;\ n \geq 1\}$ *satisfies the requirements of the Gärtner-Ellis theorem with limiting log-moment generating function*

$$\Lambda_B(\theta) \stackrel{\Delta}{=} \lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{\theta S_{1,n}^B}] \tag{16}$$

and large deviations rate function

$$\Lambda_B^*(a) \stackrel{\Delta}{=} \sup_\theta (\theta a - \Lambda_B(\theta)). \tag{17}$$

Assumption C

1. *For every* $\epsilon_1, \epsilon_2, a > 0$, *there exists* M_A *such that for all* $n \geq M_A$

$$e^{-n(\Lambda_A^{*-}(a)+\epsilon_2)} \leq \mathbf{P}[S_{1,i}^A - ia \leq \epsilon_1 n,\ i = 1, \ldots, n]. \tag{18}$$

2. *For every* $\epsilon_1, \epsilon_2, a > 0$, *there exists* M_B *such that for all* $n \geq M_B$

$$e^{-n(\Lambda_B^{*-}(a)+\epsilon_2)} \leq \mathbf{P}[S_{i,j}^B - (j-i+1)a \leq \epsilon_1 n,\ 1 \leq i \leq j \leq n], \tag{19}$$

and

$$e^{-n(\Lambda_B^{*+}(a)+\epsilon_2)} \leq \mathbf{P}[S_{i,j}^B - (j-i+1)a \geq -\epsilon_1 n,\ 1 \leq i \leq j \leq n]. \tag{20}$$

We consider external arrival and service processes that satisfy Assumptions B and C. We will show that these assumptions are satisfied by the processes resulting from the three operations mentioned above. In this way, our approach provides a *calculus of acyclic networks* since we will be able to determine the large deviations behaviour of each individual queue inductively.

Assumption B provides a LDP for the arrival and service processes. Based on these LDP's we will derive LDP's for all the processes of interest in the network. Note that only the tail probability of the external arrival processes corresponding to "many arrivals" is characterized by Assumption B. We will prove that in order to estimate probabilities of large waiting times and long delays, as we do in this paper, only such a tail probability of the aggregate arrival process in each queue of the network is needed.

Assumption C is needed in order to derive a LDP for the departure process of a G/G/1 queue. It intuitively asserts that besides the LDP for the partial sum random *variable* $S_{1,n}$, we also have a LDP for the partial sum *process* $\{S_{1,i},\ i = 1, \ldots, n\}$ for the arrivals and $\{S_{i,j},\ 1 \leq i \leq j \leq n\}$ for the service times. In other words, (18) and (19) guarantee that the partial sum process follows a path that never overshoots the straight line of slope a, in order to reach an improbable level $S_{1,n} \leq na$. A similar interpretation can be given to (20). When interarrival and service times are independent, Assumption C is a consequence of Mogulskii's theorem (see [DZ93b]). However, mild mixing conditions on the arrival and service processes suffice to guarantee Assumption C. A thorough treatment is given in [DZ93a]. In the Appendix we provide some conditions under which Assumption C is satisfied based on the results of [DZ93a].

Assumptions B and C are satisfied by processes that are used to model external arrival and services in communications networks, such as renewal processes, Markov-modulated processes and stationary processes with mild mixing conditions.

4 Large Deviations of a G/G/1 Queue

In this section, we establish a LDP for the Palm distributions of the steady-state waiting time and queue length (i.e., as these random variables are seen by a typical customer), in a G/G/1 queue with stationary arrivals and service times.

The setting is the same as in Section 3. We denote by $\{A_i,\ i \in \mathbb{Z}\}$ the stationary aggregate arrival process to the queue and we assume that it satisfies Assumption B.1. We also denote by $\{B_i,\ i \in \mathbb{Z}\}$ the stationary service process and we assume that it satisfies Assumption B.2. For this section, the independence assumption for the service times can

be relaxed. For stability purposes, we further assume $\mathbf{E}[A] > \mathbf{E}[B]$, where A (resp. B) denotes a typical interarrival (resp. service) time.

4.1 Large Deviations of the Waiting Time

Let us first characterize the steady-state waiting time, W, incurred by a typical customer. By W_n we denote the waiting time of the nth customer. The condition $\mathbf{E}[A] > \mathbf{E}[B]$ is necessary [1] for the existence and the uniqueness of a stationary process (see [Wal88]). From the Lindley equation, the waiting time of the 0th customer, at steady-state, is given by

$$W_0 = [W_{-1} + B_{-1} - A_0]^+ \triangleq \max[W_{-1} + B_{-1} - A_0, 0] = \max_{i \geq 0}[S^B_{-i-1,-1} - S^A_{-i,0}, 0]. \quad (21)$$

The intuitive meaning of this relation is the following: For a particular sample path, if i^* is the optimum i, then the customer with label $-i^* - 1$ is the one who initializes the busy period in which the 0th customer is served.

The next theorem establishes a LDP for W_0. This result is not new. The proof is almost identical with the proof in [dVW93], where a discrete time model is used, and is therefore omitted. An upper bound on the tail probability, of the steady-state waiting time, was first obtained by Kingman [Kin70].

Theorem 4.1 *The tail of the Palm distribution of the steady-state waiting time, W, in a FCFS G/G/1 queue with arrivals and service times satisfying Assumption B is characterized by*

$$\lim_{U \to \infty} \frac{1}{U} \log \mathbf{P}[W \geq U] = \theta^*, \quad (22)$$

where $\theta^* < 0$ is the smallest root of the equation

$$\Lambda_A(\theta) + \Lambda_B(-\theta) = 0. \quad (23)$$

Remarks : Intuitively, Theorem 4.1 asserts that for large enough U, we can state

$$\mathbf{P}[W \geq U] \sim e^{\theta^* U}, \qquad \text{where } \theta^* < 0 \text{ is such that} \qquad \Lambda_A(\theta^*) + \Lambda_B(-\theta^*) = 0. \quad (24)$$

Moreover, due to the strict convexity of $\Lambda_A(\cdot), \Lambda_B(\cdot)$ [2], θ^* is the unique non-zero root. Figure 2 depicts the function $\Lambda_A(\theta) + \Lambda_B(-\theta)$ and the root θ^*.

It is instructive to characterize the most likely "path" along which the large deviation of the waiting time occurs. Such a characterization can also provide an alternative proof of

[1] for sufficiency ergodicity is also needed.

[2] In [Roc70] it is proven that the convex dual of a function which satisfies Assumption A, is strictly convex on every convex subset of its domain.

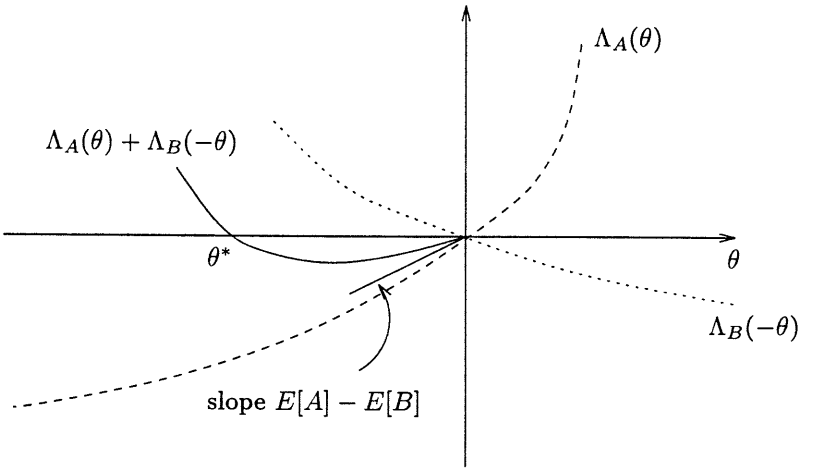

Figure 2: The root of $\Lambda_A(\theta) + \Lambda_B(-\theta) = 0$.

Thm. 4.1. Let $x_1, x_2 \in \mathbb{R}^+$, such that $x_2 - x_1 = a$. Using Eq. (21), we have

$$\begin{aligned} \mathbf{P}[W_0 \geq (i+1)a] &\geq \mathbf{P}[S^B_{-i-1,-1} - S^A_{-i,0} \geq (i+1)a] \\ &\geq \mathbf{P}[S^A_{-i,0} \leq (i+1)x_1]\mathbf{P}[S^B_{-i-1,-1} \geq (i+1)x_2] \\ &\geq e^{-(i+1)[\Lambda^{*-}_A(x_1)+\Lambda^{*+}_B(x_2)+\epsilon]}, \end{aligned} \tag{25}$$

where the last inequality makes use of Assumption B and holds for any $\epsilon > 0$ and for large i.

Setting $U = (i+1)a$, we obtain

$$\mathbf{P}[W_0 \geq U] \geq \exp\left\{-U \inf_a \frac{1}{a} \inf_{x_2 - x_1 = a} [\Lambda^{*-}_A(x_1) + \Lambda^{*+}_B(x_2)] - U\epsilon\right\}. \tag{26}$$

Let a^* be the solution to the above optimization problem. Thus, for large U, and by taking $\epsilon \to 0$ in (26), we obtain

$$\mathbf{P}[W_0 \geq U] \geq \exp\left\{-U \frac{\inf_{x_2 - x_1 = a^*} [\Lambda^{*-}_A(x_1) + \Lambda^{*+}_B(x_2)]}{a^*}\right\}. \tag{27}$$

The tightness of this bound can be proven by obtaining a matching (i.e., with the same exponent) upper bound; the proof is omitted. Let i^* be defined by the equation $i^* + 1 = U/a^*$. Let also x_1^* and x_2^* solve the optimization problem in (27). Consider a scenario where customers $(-i^* - 1), \ldots, -1$ arrive at an empirical arrival rate of $\frac{1}{x_1^*}$ and customers $-i^*, \ldots, 0$ are served with an empirical service rate of $\frac{1}{x_2^*}$. Such a scenario, which is depicted in Figure 3, has probability comparable to the right hand side of (27) and is therefore a most likely way for the large deviation of the waiting time to occur.

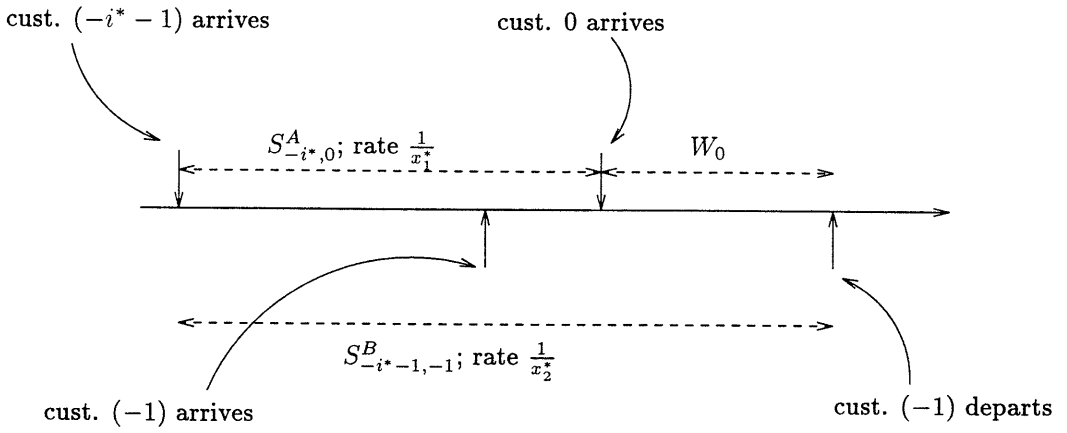

Figure 3: The optimal path for large deviations in the waiting time.

4.2 Large Deviations of the Queue Length

In this subsection, we present a LDP for the steady-state queue length in a G/G/1 queue, as seen by a typical customer (Palm distribution). To accomplish that we use the main argument used in deriving distributional laws; that is, a probabilistic relation between the waiting time and the queue length. A detailed discussion of distributional laws and their applications can be found in [BN91, BM92]. It is important to note that distributional laws have been proven there only for renewal arrival and service processes. However in the large deviations setting, we are able to relax the renewal assumption and state a result that holds even for correlated arrival and service processes.

Let us now characterize the steady-state queue length L seen by a typical customer upon arrival (this is sometimes denoted by L^- in the literature). The goal is to estimate $\mathbf{P}[L \geq n]$. Let us denote by L_{n-1} the queue length observed by the $(n-1)$st customer. As in Section 3, we assume that the process $\{(L_n, W_n);\ n \in \mathbb{Z}\}$ is stationary. The main idea, in order to establish a relation between the waiting time and the queue length, is to look backwards in time from the arrival epoch of the $(n-1)$st customer. Figure 4 depicts the situation. We

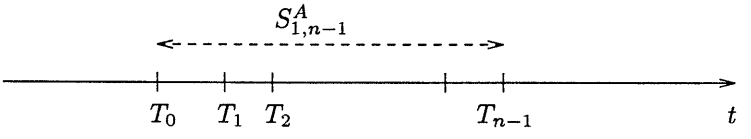

Figure 4: The system at time T_{n-1}.

denote with $T_0, T_1, \ldots$ the arrival epochs of customers $0, 1, \ldots$, respectively. Recall that W_n and B_n denote the waiting and the service time of the nth customer, respectively.

The main observation is the following: In order for the queue length right after T_{n-1} to

be at least n, the 0th customer should be in the system at time T_{n-1}. Namely,

$$\mathbf{P}[L_{n-1} \geq n] = \mathbf{P}[W_0 + B_0 \geq S^A_{1,n-1}] \tag{28}$$

and by using (21) we obtain

$$\begin{aligned}\mathbf{P}[L_{n-1} \geq n] &= \mathbf{P}[\max_{i\geq 0}[S^B_{-i-1,0} - S^A_{-i,0}, -S^A_{1,n-1}] \geq 0] = \\ &= \mathbf{P}[\max_{i\geq -1}[S^B_{-i-1,0} - S^A_{-i,n-1}] \geq 0].\end{aligned} \tag{29}$$

The next theorem establishes a LDP for L_{n-1}. We will need a technical lemma which we prove next. The notational convention $S^X_{i,j} \triangleq 0$; $i > j$, is used for the partial sums of the random sequence $\{X_i;\ i \in \mathbb{Z}\}$.

Lemma 4.2 *Under Assumption B, and for $\theta < 0$, satisfying $\Lambda_A(\theta) + \Lambda_B(-\theta) < 0$, it holds*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{-\theta \max_{i\geq -1}[S^B_{-i-1,0} - S^A_{-i,n-1}]}] \leq \Lambda_A(\theta). \tag{30}$$

Proof : Fix some $\theta < 0$ satisfying $\Lambda_A(\theta) + \Lambda_B(-\theta) < 0$ and some $\epsilon > 0$ such that $\Lambda_A(\theta) + \Lambda_B(-\theta) + 2\epsilon < 0$. Note that the existence of such a θ is guaranteed by the condition $\mathbf{E}[A] > \mathbf{E}[B]$ (see Figure 2) [3]. Notice now that

$$\mathbf{E}[e^{-\theta \max_{i\geq -1}[S^B_{-i-1,0} - S^A_{-i,n-1}]}] \leq \sum_{i\geq -1} \mathbf{E}[e^{-\theta S^B_{-i-1,0}}]\mathbf{E}[e^{\theta S^A_{-i,n-1}}].$$

From (14) and (16) it can be seen that there exists $j > 0$ such that for all $i > j$ and all $n \geq 0$ it holds

$$\mathbf{E}[e^{\theta S^A_{-i,n-1}}] \leq e^{(n+i)(\Lambda_A(\theta)+\epsilon)} \qquad \text{and} \qquad \mathbf{E}[e^{-\theta S^B_{-i-1,0}}] \leq e^{(i+2)(\Lambda_B(-\theta)+\epsilon)}.$$

We then have

$$\begin{aligned}&\mathbf{E}[e^{-\theta \max_{i\geq -1}[S^B_{-i-1,0} - S^A_{-i,n-1}]}] \leq \\ &\leq \sum_{i=-1}^{j} \mathbf{E}[e^{-\theta S^B_{-i-1,0}}]\mathbf{E}[e^{\theta S^A_{-i,n-1}}] + e^{n(\Lambda_A(\theta)+\epsilon)} e^{2(\Lambda_B(-\theta)+\epsilon)} \sum_{i>j} e^{i(\Lambda_A(\theta)+\Lambda_B(-\theta)+2\epsilon)} \\ &\leq K(\theta, j, \epsilon) e^{n(\Lambda_A(\theta)+\epsilon)},\end{aligned} \tag{31}$$

where $K(\theta, j, \epsilon)$ is some constant depending on θ, j and ϵ but not on n. To see that, notice that with θ and ϵ chosen as above, the infinite geometric series in (31) converges to a constant independent of n. Also notice that the finite sum from -1 to j in (31) can be upper bounded by a constant independent of n; this is because the only term that depends

[3] To see that note that for small θ we have $\Lambda_A(\theta) + \Lambda_B(-\theta) = \theta(\mathbf{E}[B] - \mathbf{E}[A]) + o(\theta)$.

on n is $\mathbf{E}[e^{\theta S^A_{-i,n-1}}]$ which is bounded above by 1 since $\theta < 0$. From Eq. (31) we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{-\theta \max_{i\geq -1}[S^B_{-i-1,0} - S^A_{-i,n-1}]}] \leq \Lambda_A(\theta) + \epsilon. \tag{32}$$

Since this is true for all small enough $\epsilon > 0$, the result follows.

■

Theorem 4.3 *The tail of the Palm distribution of the steady-state queue length, L, in a FCFS G/G/1 queue with arrivals and service times satisfying Assumption B is characterized by*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[L \geq n] = \Lambda_A(\theta^*), \tag{33}$$

where $\theta^* < 0$ is the smallest root of the equation

$$\Lambda_A(\theta) + \Lambda_B(-\theta) = 0. \tag{34}$$

Proof : Due to stationarity, it suffices to characterize the tail distribution of L_{n-1}. For an upper bound define

$$G_n \stackrel{\Delta}{=} \max_{i\geq -1}[S^B_{-i-1,0} - S^A_{-i,n-1}]. \tag{35}$$

Using the Markov inequality, we obtain

$$\mathbf{P}[L_{n-1} \geq n] = \mathbf{P}[G_n \geq 0] \leq \mathbf{E}[e^{-\theta G_n}],$$

for $\theta < 0$. Taking the limit as $n \to \infty$, using Lemma 4.2, and optimizing over θ to get the best bound we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[L_{n-1} \geq n] \leq \inf_{\{\theta |\ \Lambda_A(\theta)+\Lambda_B(-\theta)<0\}} [\Lambda_A(\theta)] = \Lambda_A(\theta^*), \tag{36}$$

where the last equality is justified by Figure 2.

For a lower bound, set $i = \delta n$ for $\delta \geq 0$ (δn is assumed integer), and notice that

$$\begin{aligned}
\mathbf{P}[L_{n-1} \geq n] &= \mathbf{P}[G_n \geq 0] \\
&= \mathbf{P}[\max_{\delta\geq 0}[S^B_{-\delta n-1,0} - S^A_{-\delta n,n-1}, -S^A_{1,n-1}] \geq 0] \\
&\geq \sup_{\delta\geq 0} \mathbf{P}[S^B_{-\delta n-1,0} - S^A_{-\delta n,n-1} \geq 0].
\end{aligned}$$

The limiting log-moment generating function of $S^B_{-\delta n-1,0} - S^A_{-\delta n,n-1}$ is

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{-\theta(S^B_{-\delta n-1,0} - S^A_{-\delta n,n-1})}] = \delta\Lambda_B(-\theta) + (1+\delta)\Lambda_A(\theta)$$

and by using Assumption B we obtain

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}[L_{n-1} \geq n] &\geq \sup_{\delta\geq 0}(-\sup_{\theta}[-\delta(\Lambda_A^-(\theta)+\Lambda_B^+(-\theta))-\Lambda_A^-(\theta)]) \\
&= \sup_{\delta\geq 0}\inf_{\theta}[\delta(\Lambda_A^-(\theta)+\Lambda_B^+(-\theta))+\Lambda_A^-(\theta)] \\
&= \inf_{\{\theta|\ \Lambda_A^-(\theta)+\Lambda_B^+(-\theta)<0\}}[\Lambda_A^-(\theta)] \\
&= \Lambda_A^-(\theta^*) = \Lambda_A(\theta^*), \qquad (37)
\end{aligned}$$

where the second equality follows by dualizing the constraint $\Lambda_A^-(\theta)+\Lambda_B^+(-\theta) < 0$. The lower bound in (37) along with (36) proves (33).

■

Remark : Intuitively, Theorem 4.3 asserts that for large enough n, we can state

$$\mathbf{P}[L \geq n] \sim e^{n\Lambda_A(\theta^*)}, \qquad \text{where } \theta^* < 0 \text{ such that} \qquad \Lambda_A(\theta^*)+\Lambda_B(-\theta^*) = 0. \qquad (38)$$

5 The Departure Process of a G/GI/1 queue

In this section we obtain a LDP for the process resulting from the passing-through-a-queue operation of our network model. That is, we establish a LDP for the steady-state departure process of a G/GI/1 queue, as seen by a typical departing customer. We denote by D_i, $i \in \mathbb{Z}$, the inter-departure time of the ith customer (the interval between the departure epochs of the $(i-1)$st and the ith customer). As in Section 3 we assume that the interarrival times process $\{A_i,\ i \in \mathbb{Z}\}$ is stationary, and A_i are possibly dependent random variables. The service times B_i are independent and identically distributed (iid) random variables. The arrival and service processes are also assumed to satisfy Assumptions B and C. As explained in Section 3, we will prove that the departure process satisfies Assumptions B and C when the arrival and service processes do.

We denote by $S_{1,n}^D \triangleq \sum_{i=1}^n D_i$, the partial sum of the departure process. The objective of this section is to prove a LDP for $S_{1,n}^D$. The inter-departure times can be expressed as follows

$$D_i = B_i + I_i, \qquad (39)$$

where B_i denotes the service time of the ith customer and I_i the idling period of the system that ended with the arrival of the ith customer ($I_i = 0$ if the ith customer finds the system busy upon arrival). By using the Lindley equation one can obtain an expression for I_i and after some algebra derive an expression for $S_{1,n}^D$ in terms of the partial sums for the arrival and the service process. Using such an expression one can prove a LDP for $S_{1,n}^D$. For a detailed exposition of this approach the reader is referred to [Pas]. In this paper we follow

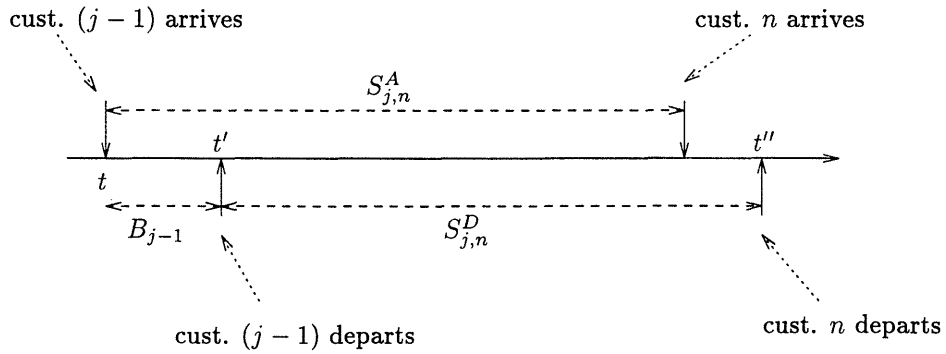

Figure 5: Deriving an upper bound on $\mathbf{P}[S^D_{1,n} \leq na]$.

a more intuitive approach. We derive an upper bound and a matching lower bound on $\mathbf{P}[S^D_{1,n} \leq na]$ based on sample path arguments. To that effect, we explicitly characterize the most likely path leading to the large deviation of the departure process. The next proposition establishes an upper bound for the tail probability of $S^D_{1,n}$.

Proposition 5.1 (Upper Bound) *Under Assumption B, the partial sum $S^D_{1,n}$ of the departure process of a G/GI/1 queue under FCFS satisfies*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S^D_{1,n} \leq na] \leq -\Lambda^{*-}_D(a), \tag{40}$$

where

$$\Lambda^{*-}_D(a) \stackrel{\Delta}{=} \Lambda^{*-}_B(a) + \Lambda^{*-}_\Gamma(a) \tag{41}$$

and

$$\Lambda^{*-}_\Gamma(a) \stackrel{\Delta}{=} \sup_{\{\theta | \Lambda_A(\theta)+\Lambda_B(-\theta)<0\}} [\theta a - \Lambda^-_A(\theta)]. \tag{42}$$

Proof : Since $D_i \geq B_i$ for all i we obtain

$$S^D_{1,n} \geq S^B_{1,n}. \tag{43}$$

Consider some $j \leq 1$ and let $(j-1)$ be the customer who initializes the busy period in which the 0th customer is served. Let t be the time that the $(j-1)$st customer arrived, t' the time that the $(j-1)$st customer departed, and t'' the time that the nth customer departed. Figure 5 depicts the situation. Note that

$$B_{j-1} + S^D_{j,n} \geq S^A_{j,n}. \tag{44}$$

Since the system is busy from the arrival of the $(j-1)$st customer until the departure of customer 0, we have

$$S^D_{j,0} = S^B_{j,0}. \tag{45}$$

Therefore, from (45) and (44) we have

$$S^D_{1,n} = S^D_{j,n} - S^D_{j,0} \geq S^A_{j,n} - B_{j-1} - S^B_{j,0} = S^A_{j,n} - S^B_{j-1,0}. \tag{46}$$

Now, from (43) and (46) we obtain

$$\begin{aligned} \mathbf{P}[S^D_{1,n} \leq na] &\leq \mathbf{P}[S^B_{1,n} \leq na,\ \exists\ j \leq 1 \text{ s.t. } S^A_{j,n} - S^B_{j-1,0} \leq na] \\ &= \mathbf{P}[S^B_{1,n} \leq na]\mathbf{P}[\min_{j\leq 1}[S^A_{j,n} - S^B_{j-1,0}] \leq na], \end{aligned} \tag{47}$$

since the service times B_i are assumed to be independent and independent of the arrival process. Notice that

$$\min_{j\leq 1}[S^A_{j,n} - S^B_{j-1,0}] = A_n - \max_{j\leq 1}[S^B_{j-1,0} - S^A_{j,n-1}].$$

Since the moment generating function of A_n is independent of n, we use Lemma 4.2 to obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{\theta \min_{j\leq 1}[S^A_{j,n} - S^B_{j-1,0}]}] \leq \Lambda_A(\theta), \tag{48}$$

for $\theta < 0$, satisfying $\Lambda_A(\theta) + \Lambda_B(-\theta) < 0$.

Using Markov's inequality we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[\min_{j\leq 1}[S^A_{j,n} - S^B_{j-1,0}] \leq na] \leq \Lambda_A(\theta) - \theta a.$$

Optimizing over θ to obtain the tightest bound we finally find (note that for $\theta < 0$ we have $\Lambda^-_A(\theta) = \Lambda_A(\theta)$)

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[\min_{j\leq 1}[S^A_{j,n} - S^B_{j-1,0}] \leq na] \leq - \sup_{\{\theta | \Lambda_A(\theta)+\Lambda_B(-\theta)<0\}} [\theta a - \Lambda^-_A(\theta)]. \tag{49}$$

Moreover from Assumption B we can assert that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S^B_{1,n} \leq na] \leq -\Lambda^{*-}_B(a). \tag{50}$$

Combining Eqs. (50) and (49) along with Eq. (47) we obtain (40).

■

Obtaining a lower bound on the tail probability of $S^D_{1,n}$ is much more involved. Assump-

tion B which provides a LDP for the partial sums $S^A_{1,n}, S^B_{1,n}$ of the interarrival and service times is not sufficient. Assumption C which provides a LDP for the partial sum processes $\{S^A_{1,j},\ j = 1,\dots,n\}$ and $\{S^B_{i,j},\ 1 \le i \le j \le n\}$, is required. In the next proposition we derive a lower bound on the tail probability of $S^D_{1,n}$ and we prove that the departure process $\{S^D_{1,i},\ i = 1,\dots,n\}$ satisfies Assumption C.

Proposition 5.2 (Lower Bound) *Under Assumptions B and C, the partial sum $S^D_{1,n}$ of the departure process of a G/GI/1 queue under FCFS satisfies*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S^D_{1,n} \le na] \ge -\Lambda^{*-}_D(a). \tag{51}$$

Moreover, the departure process $\{S^D_{1,i},\ i = 1,\dots,n\}$ satisfies Assumption C.

Proof : We distinguish two cases: $a < \mathbf{E}[B]$ and $a \ge \mathbf{E}[B]$, where $\mathbf{E}[B]$ denotes the mean service time. For the case $a < \mathbf{E}[B]$, we fix $\epsilon_1, \epsilon_2 > 0$, $\zeta \ge 0$, $y_1, y_2 \ge 0$ such that $y_1 - y_2 = a$ and $\frac{y_1}{1+\zeta} \ge a$. Consider the set of all sample paths that satisfy

$$(k+1-j)a - \epsilon_1 n \le S^B_{j,k} \le (k+1-j)a + \epsilon_1 n, \qquad 1 \le j \le k \le n, \tag{52}$$

$$S^A_{-\zeta n,k} \le (\zeta n + k - 1)\frac{y_1}{1+\zeta} + \epsilon_1 n, \qquad k = 1,\dots,n, \tag{53}$$

and

$$S^B_{-\zeta n-1,0} \ge n y_2 + 2n\epsilon_1. \tag{54}$$

We state the following lemma the proof of which is deferred until the end of the current proof.

Lemma 5.3 *For any sample path that satisfies (52), (53) and (54) we have*

$$D_k = B_k, \qquad k = 1,\dots,n. \tag{55}$$

As a consequence it is clear from (52) that within this set of sample paths we have $S^D_{1,i} \le ia + \epsilon_1 n,\ i = 1,\dots,n$. Therefore,

$$\begin{aligned}
&\mathbf{P}[S^D_{1,i} \le ia + \epsilon_1 n,\ i = 1,\dots,n] \ge \\
&\quad \mathbf{P}[(k+1-j)a - \epsilon_1 n \le S^B_{j,k} \le (k+1-j)a + \epsilon_1 n,\ 1 \le j \le k \le n] \times \\
&\qquad \sup_{\{\zeta \ge 0 \mid \frac{y_1}{1+\zeta} \ge a\}} \ \sup_{y_1 - y_2 = a} \mathbf{P}[S^A_{-\zeta n,k} \le (\zeta n + k - 1)\frac{y_1}{1+\zeta} + \epsilon_1 n,\ k = 1,\dots,n] \times \\
&\qquad\qquad \mathbf{P}[S^B_{-\zeta n-1,0} \ge n y_2 + 2n\epsilon_1]
\end{aligned}$$

$$\geq \sup_{\{\zeta\geq 0|\frac{y_1}{1+\zeta}\geq a\}} \sup_{y_1-y_2=a} \exp\Big\{-n(\Lambda_B^{*-}(a)+\epsilon') - n\left[\Lambda_A^{*-}\left(\tfrac{y_1}{1+\zeta}\right)(1+\zeta)+\epsilon''\right]$$

$$- n\left[\Lambda_B^{*+}\left(\tfrac{y_2+2\epsilon_1}{\zeta}\right)\zeta+\epsilon'''\right]\Big\}, \tag{56}$$

where the last inequality holds for large n and is obtained by applying Assumption C to the arrival and service processes. We can now choose appropriate ϵ', ϵ'' and ϵ''' such that for sufficiently large n and given ϵ_2 we have

$$\mathbf{P}[S_{1,i}^D \leq ia+\epsilon_1 n,\ i=1,\dots,n] \geq \sup_{\{\zeta\geq 0|\frac{y_1}{1+\zeta}\geq a\}} \sup_{y_1-y_2=a} \exp\Big\{-n\left[\Lambda_B^{*-}(a)+\right.$$

$$\left.\Lambda_A^{*-}\left(\tfrac{y_1}{1+\zeta}\right)(1+\zeta)+\Lambda_B^{*+}\left(\tfrac{y_2}{\zeta}\right)\zeta+\epsilon_2\right]\Big\}. \tag{57}$$

Now note that we can remove the constraint $\frac{y_1}{1+\zeta}\geq a$ from the optimization in (57), since we are in the region $a<\mathbf{E}[B]$. To see that consider a choice of y_1, y_2 such that $y_1-y_2=a$ and $a>\frac{y_1}{1+\zeta}$. This implies $\frac{y_2}{\zeta}<a$. Let us now increase y_1 so that $\frac{y_1}{1+\zeta}=a$ and $\frac{y_2}{\zeta}=a<\mathbf{E}[B]$. Then, $\Lambda_A^{*-}(\frac{y_1}{1+\zeta})$ decreases while $\Lambda_B^{*+}(\frac{y_2}{\zeta})$ stays at zero. Hence, values of y_1 such that $a>\frac{y_1}{1+\zeta}$ are dominated by $y_1=(1+\zeta)a$ which proves that the constraint $\frac{y_1}{1+\zeta}\geq a$ can be removed.

We now use convex analysis to prove that $\Lambda_\Gamma^{*-}(a)$ as defined in Eq. (42) is equal to

$$-\sup_{\zeta\geq 0}\sup_{y_1-y_2=a}\left\{-(1+\zeta)\Lambda_A^{*-}\left(\tfrac{y_1}{1+\zeta}\right)-\zeta\Lambda_B^{*+}\left(\tfrac{y_2}{\zeta}\right)\right\}$$

thus, proving that the lower bound in (57) (taking $\epsilon_2\to 0$) matches the upper bound obtained in Proposition 5.1. Dualizing the constraint $\Lambda_A(\theta)+\Lambda_B(-\theta)<0$ we obtain (note that $\Lambda_A(\theta)+\Lambda_B(-\theta)<0$ if and only if $\Lambda_A^-(\theta)+\Lambda_B^+(-\theta)<0$)

$$\begin{aligned}
-\Lambda_\Gamma^{*-}(a) &= -\sup_{\{\theta|\Lambda_A(\theta)+\Lambda_B(-\theta)<0\}}[\theta a-\Lambda_A^-(\theta)] \\
&= \inf_{\{\theta|\Lambda_A(\theta)+\Lambda_B(-\theta)<0\}}[-\theta a+\Lambda_A^-(\theta)] \\
&= \sup_{\zeta\geq 0}\left\{-\sup_\theta[\theta a-(1+\zeta)\Lambda_A^-(\theta)-\zeta\Lambda_B^+(-\theta)]\right\} \\
&= \sup_{\zeta\geq 0}\left\{-\inf_{y_1-y_2=a}\left[(1+\zeta)\Lambda_A^{*-}\left(\tfrac{y_1}{1+\zeta}\right)+\zeta\Lambda_B^{*+}\left(\tfrac{y_2}{\zeta}\right)\right]\right\} \\
&= \sup_{\zeta\geq 0}\sup_{y_1-y_2=a}\left[-(1+\zeta)\Lambda_A^{*-}\left(\tfrac{y_1}{1+\zeta}\right)-\zeta\Lambda_B^{*+}\left(\tfrac{y_2}{\zeta}\right)\right].
\end{aligned} \tag{58}$$

To see that, note that for convex functions f, f_1, f_2 and for a scalar $c\geq 0$, it holds $(cf)^*(x^*)=cf^*(x^*/c)$, and $(f_1+f_2)^*(x^*)=\inf_{x_1^*+x_2^*=x^*}[f_1^*(x_1^*)+f_2^*(x_2^*)]$ (see [Roc70] for details).

In summary, for the region $a<\mathbf{E}[B]$ we have verified that Assumption C holds for the

departure process, i.e.,

$$\mathbf{P}[S^D_{1,i} \leq ia + \epsilon_1 n,\ i = 1, \ldots, n] \geq e^{-n(\Lambda^{*-}_D(a)+\epsilon_2)}. \tag{59}$$

By taking $\epsilon_1, \epsilon_2 \to 0$ and since $\mathbf{P}[S^D_{1,n} \leq na]$ is clearly larger than the probability in (59), (51) is verified for the same region.

We now consider the region $a \geq \mathbf{E}[B]$. Fix $\epsilon_1, \epsilon_2 > 0$, $\zeta \geq 0$ and $y_1, y_2 \geq 0$ such that $y_1 - y_2 = a$ and $\frac{y_1}{1+\zeta} \geq a$. Consider the set of all sample paths that satisfy

$$S^B_{j,k} \leq (k + 1 - j)a + \epsilon_1 n, \qquad 1 \leq j \leq k \leq n, \tag{60}$$

$$S^A_{-\zeta n,k} \leq (\zeta n + k - 1)\frac{y_1}{1+\zeta} + \epsilon_1 n, \qquad k = 1, \ldots, n, \tag{61}$$

and

$$S^B_{-\zeta n-1,0} \geq n y_2 - \epsilon_1 n. \tag{62}$$

We state the following lemma the proof of which is also deferred until the end of the current proof.

Lemma 5.4 *For any sample path that satisfies (60), (61) and (62) we have*

$$S^D_{1,k} \leq ka + 3\epsilon_1 n, \qquad k = 1, \ldots, n. \tag{63}$$

Therefore,

$$\begin{aligned}
&\mathbf{P}[S^D_{1,k} \leq ka + 3\epsilon_1 n,\ k = 1, \ldots, n] \geq \\
&\quad \mathbf{P}[S^B_{j,k} \leq (k + 1 - j)a + \epsilon_1 n,\ 1 \leq j \leq k \leq n] \times \\
&\qquad \sup_{\{\zeta \geq 0 | \frac{y_1}{1+\zeta} \geq a\}} \sup_{y_1 - y_2 = a} \mathbf{P}[S^A_{-\zeta n,k} \leq (\zeta n + k - 1)\tfrac{y_1}{1+\zeta} + \epsilon_1 n,\ k = 1, \ldots, n] \times \\
&\qquad\quad \mathbf{P}[S^B_{-\zeta n-1,0} \geq n y_2 - \epsilon_1 n].
\end{aligned} \tag{64}$$

Since we are in the region $a \geq \mathbf{E}[B]$, the first term in the right hand side of (64) is $O(1)$. By similar calculations as the ones performed in the case $a < \mathbf{E}[B]$ (see Eq. (57)) it can be verified that

$$\begin{aligned}
\mathbf{P}[S^D_{1,k} \leq ka + 3\epsilon_1 n,\ k = 1, \ldots, n] \geq \sup_{\{\zeta \geq 0 | \frac{y_1}{1+\zeta} \geq a\}} \sup_{y_1 - y_2 = a} \exp\Big\{-n\Big[\Lambda^{*-}_A\Big(\tfrac{y_1}{1+\zeta}\Big)(1+\zeta) + \\
+\Lambda^{*+}_B\Big(\tfrac{y_2}{\zeta}\Big)\zeta + \epsilon_2\Big]\Big\}.
\end{aligned} \tag{65}$$

We now argue that the constraint $\frac{y_1}{1+\zeta} \geq a$ can be removed from the optimization in (65). Consider a choice of $y_1 = \tilde{y}_1$, $y_2 = \tilde{y}_2$ and $\zeta = \tilde{\zeta}$ such that $\tilde{y}_1 - \tilde{y}_2 = a$ and $\frac{\tilde{y}_1}{1+\tilde{\zeta}} < a$. Let us now consider the subset of sample paths with $\zeta = 0$, $y_1 = a$ and $y_2 = 0$ from those satisfying (60), (61) and (62). It is easy to see that the probability of this subset is $e^{-n\Lambda_A^{*-}(a)}$. Now note that since $\frac{\tilde{y}_1}{1+\tilde{\zeta}} < a$ we have

$$\exp\{-n\Lambda_A^{*-}(a)\} > \exp\left\{-n\left[\Lambda_A^{*-}\left(\frac{\tilde{y}_1}{1+\tilde{\zeta}}\right)(1+\tilde{\zeta}) + \Lambda_B^{*+}\left(\frac{\tilde{y}_2}{\tilde{\zeta}}\right)\tilde{\zeta}\right]\right\}.$$

This shows that there exist choices of y_1, y_2 and ζ satisfying $\frac{y_1}{1+\zeta} \geq a$ that have a better exponent. Hence, the constraint $\frac{y_1}{1+\zeta} \geq a$ can indeed be removed.

Thus, for the region $a \geq \mathbf{E}[B]$ also, Assumption C holds for the departure process, i.e., for all $\epsilon_1, \epsilon_2 > 0$ and for large enough n we have

$$\mathbf{P}[S_{1,k}^D \leq ka + 3\epsilon_1 n,\ k = 1, \ldots, n] \geq e^{-n(\Lambda_\Gamma^{*-}(a)+\epsilon_2)} = e^{-n(\Lambda_D^{*-}(a)+\epsilon_2)}. \tag{66}$$

Note that when $a \geq \mathbf{E}[B]$ we have $\Lambda_\Gamma^{*-}(a) = \Lambda_D^{*-}(a)$. By taking $\epsilon_1, \epsilon_2 \to 0$ and since $\mathbf{P}[S_{1,n}^D \leq na]$ is clearly larger than the probability in (66), (51) is also verified for the same region.

■

Proof of Lemma 5.3: Note that for $k = 1, \ldots, n$ from (53) we obtain

$$\begin{aligned} S_{-\zeta n,k}^A &\leq (\zeta n + k - 1)\frac{y_1}{1+\zeta} + \epsilon_1 n \\ &\leq (ny_2 + 2n\epsilon_1) + (a(k-1) - \epsilon_1 n) \\ &\leq S_{-\zeta n-1,0}^B + S_{1,k-1}^B = S_{-\zeta n-1,k-1}^B, \end{aligned} \tag{67}$$

where the second inequality holds because the two sides are equal at $k = n+1$ and because $\frac{y_1}{1+\zeta} \geq a$. The third inequality is justified by (52) and (54).

Let t be the arrival time of customer $-\zeta n - 1$. Then customer k arrives at time $t + S_{-\zeta n,k}^A$. The service time of customer $k-1$ can end no earlier than $t + S_{-\zeta n-1,k-1}^B \geq t + S_{-\zeta n,k}^A$. Hence, customer k finds a busy system upon arrival and therefore $D_k = B_k$, $k = 1, \ldots, n$.

■

The above proof indicates the most likely path along which the large deviation of $S_{1,n}^D$ occurs when $a < \mathbf{E}[B]$. Let ζ^*, y_1^* and y_2^* solve the optimization problem in (57). The large deviation in $S_{1,n}^D$ occurs by

- Maintaining an empirical arrival rate of $\frac{1+\zeta^*}{y_1^*}$ from the arrival of customer $-\zeta^* n - 1$, until the departure of the nth customer, and an empirical service rate of $\frac{\zeta^*}{y_2^*}$ from the arrival of customer $-\zeta^* n - 1$, until the departure of the 0th customer, and by

- Maintaining an empirical service rate of $1/a$ from the departure of the 0th customer until the departure of the nth customer.

Figure 6 illustrates the situation.

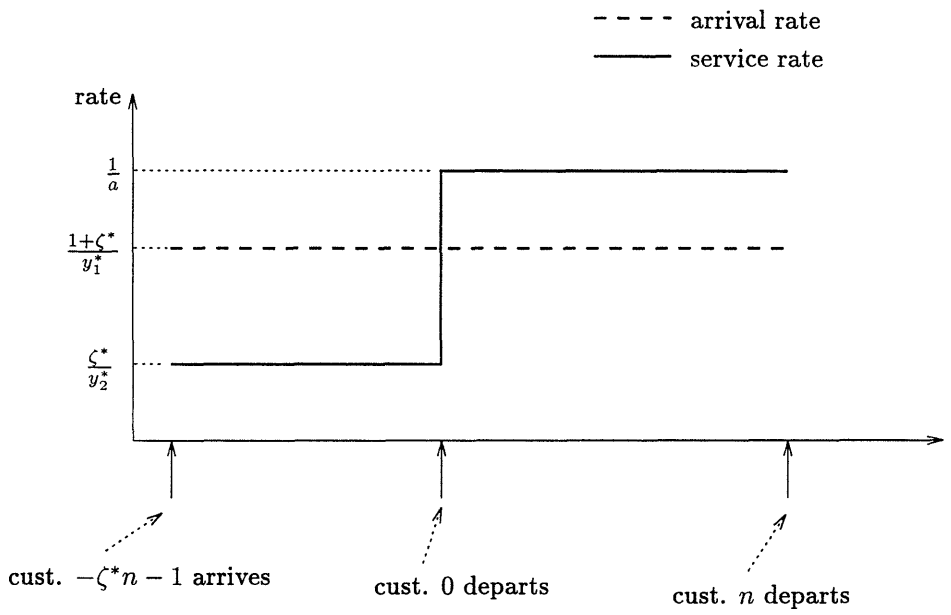

Figure 6: The most likely path for large deviations of $S^D_{1,n}$, when $a < \mathbf{E}[B]$.

Proof of Lemma 5.4: Note that for $k = 1, \ldots, n$ from (61) and (62) we obtain

$$S^A_{-\zeta n,k} \leq (ny_2 - \epsilon_1 n) + ((k-1)a + 2\epsilon_1 n) \leq (k-1)a + 2\epsilon_1 n + S^B_{-\zeta n-1,0}, \tag{68}$$

by using the argument used in Eq. (67). Let t be the arrival time of customer $-\zeta n - 1$. Then customer k arrives at time $t + S^A_{-\zeta n,k}$. We distinguish two cases. In case 1, customer k finds a busy system upon arrival in which case $D_k = B_k$. In case 2, customer k finds an empty system upon arrival. Then it departs at time t' where

$$t' = t + S^A_{-\zeta n,k} + B_k \leq ka + 3\epsilon_1 n + t + S^B_{-\zeta n-1,0}, \tag{69}$$

by using (60) and (68). Let t'' the departure time of the 0th customer. Clearly, $t'' \geq t + S^B_{-\zeta n-1,0}$, which along with (69) implies that $t' - t'' \leq ka + 3\epsilon_1 n$. But, according to their definition $t' - t'' = S^D_{1,k}$.

■

The above proof implies that the most likely path along which the deviation of $S^D_{1,n}$ occurs when $a \geq \mathbf{E}[B]$ is the following: the first departures out of customers $1, \ldots, n$ will occur at a rate of $\frac{1}{\mathbf{E}[B]} \geq \frac{1}{a}$ until the system becomes empty. Subsequent departures will

occur according to the arrival rate $\frac{1+\zeta^*}{y_1^*}$. In any case, $S_{1,k}^D \leq ka$ will hold, but the departure process will not follow a straight line scenario as it was the case in the region $a < \mathbf{E}[B]$.

Combining Propositions 5.1 and 5.2 we obtain the following theorem.

Theorem 5.5 *Under Assumptions B and C, the partial sum $S_{1,n}^D$ of the departure process of a G/GI/1 queue under FCFS satisfies*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_{1,n}^D \leq na] = -\Lambda_D^{*-}(a), \tag{70}$$

where

$$\Lambda_D^{*-}(a) = \Lambda_B^{*-}(a) + \Lambda_\Gamma^{*-}(a)$$

and

$$\Lambda_\Gamma^{*-}(a) = \sup_{\{\theta | \Lambda_A(\theta)+\Lambda_B(-\theta)<0\}} [\theta a - \Lambda_A^-(\theta)].$$

We now argue that the passing-through-a-queue operation preserves Assumption B. Proposition 5.2 establishes that it preserves Assumption C. To see that the departure process satisfies Assumption B, notice that we have proven a LDP for the departure process with large deviations rate function expressed as a function of the large deviations rate function of the arrival and service processes.

Throughout this section we have assumed that the service times B_i are iid. A close examination of the proofs of Propositions 5.1 and 5.2, suggests that a weaker condition is sufficient for our purposes. Namely, we only need the random variables $S_{j,0}^B$ and $S_{1,n}^B$ to be approximately independent for every $j \leq 0$, as $n \to \infty$. A mixing condition of the type $\mathbf{E}[e^{\theta S_{j,0}^B} e^{\theta S_{1,n}^B}] = \mathbf{E}[e^{\theta S_{j,0}^B}]\mathbf{E}[e^{\theta S_{1,n}^B}]e^{n\epsilon(n)}$ for every $j \leq 0$, where $\lim_{n\to\infty} \epsilon(n) = 0$, is sufficient.

An alternative expression for $\Lambda_D^{*-}(\cdot)$ which is a consequence of the defining Eq. (41) is

$$\Lambda_D^{*-}(a) = \Lambda_B^{*-}(a) + \Lambda_\Gamma^{*-}(a) = \begin{cases} \Lambda_B^{*-}(a) + \Lambda_A^{*-}(a) & \text{if } a \geq \Lambda_A'(\theta^*) \\ \Lambda_B^{*-}(a) + \theta^* a - \Lambda_A(\theta^*) & \text{if } a < \Lambda_A'(\theta^*) \end{cases} \tag{71}$$

where θ^* is defined in the statement of Thm. 4.1 and $\Lambda_A'(x)$ denotes the derivative of $\Lambda_A(\cdot)$ evaluated at x. To see that consult Figure 2 and notice that the first branch of Eq. (71) corresponds to the region of a where the constraint $\Lambda_A(\theta) + \Lambda_B(-\theta) < 0$ is not tight, and the second branch to the region of a where this constraint is tight.

To obtain the limiting log-moment generating function for the partial sum of the departure process, we take the convex dual of $\Lambda_D^{*-}(\cdot)$ in (71). Using the duality correspondences proven in [Roc70] we obtain the following corollary.

Corollary 5.6 *Under Assumptions B and C we have*

$$\Lambda_D^-(\theta) = \begin{cases} \inf_{\theta_1+\theta_2=\theta}\{\Lambda_B^-(\theta_1) + \Lambda_A^-(\theta_2)\} & \text{if } \theta \geq \hat{\theta} \\ \Lambda_B^-(\theta - \theta^*) + \Lambda_A(\theta^*) & \text{if } \theta < \hat{\theta} \end{cases} \tag{72}$$

where

$$\hat{\theta} \triangleq \frac{d}{da}\left[\Lambda_B^{*-}(a) + \Lambda_A^{*-}(a)\right]_{a=\Lambda_A'(\theta^*)}. \tag{73}$$

It is instructive to determine the fluctuations of the queue length that lead to a large deviation in the departure process. Let ζ^* solve the optimization problem in (57) or (65). Let t be the arrival time of customer $-\zeta^* n - 1$. The 0th customer arrives at $t + S^A_{-\zeta^* n,0}$ and departs no earlier than $t + S^B_{-\zeta^* n-1,0}$. Thus, for the waiting time of customer 0 holds

$$W_0 \geq t + S^B_{-\zeta^* n-1,0} - t - S^A_{-\zeta^* n,0} = S^B_{-\zeta^* n-1,0} - S^A_{-\zeta^* n,0} \triangleq \tilde{W}_0. \tag{74}$$

A close examination of the proofs of Propositions 5.1 and 5.2 suggests that $\Lambda_\Gamma^{*-}(\cdot)$ is the large deviations rate function of the process

$$\{S^A_{-\zeta^* n,k} - S^B_{-\zeta^* n-1,0},\ k = 1, \ldots, n\} \equiv \{S^A_{1,k} - \tilde{W}_0,\ k = 1, \ldots, n\}. \tag{75}$$

From the above discussion and Eq. (71) we conclude that depending on the value of a, we can distinguish two cases for the large deviation in the departure process to occur.

$a \geq \Lambda_A'(\theta^*)$: In this region, $\Lambda_\Gamma^{*-}(a) = \Lambda_A^{*-}(a)$ and from Eq. (75) it is clear that the most likely way for the large deviation in the departure process to occur is the 0th customer to incur $O(1)$ waiting time, which implies that it finds a queue length of $O(1)$ upon arrival.

$a < \Lambda_A'(\theta^*)$: In this region, $\Lambda_\Gamma^{*-}(a) = \theta^* a - \Lambda_A(\theta^*)$ and from Eq. (75) it is clear that the most likely way for the large deviation in the departure process to occur is the 0th customer to incur a large waiting time (recall from Thm. 4.1 that the large deviations rate function for the waiting time is linear with slope θ^*).

Hence, taking also into account Fig. 6 we can infer for the queue length the cases depicted in Figure 7. In Region 2 and in contrast with Region 1, the queue builds up to lead to a large deviation in the departure process. As we have already discussed, whether the queue empties before the departure of customer n depends on whether $a < \mathbf{E}[B]$ or $a \geq \mathbf{E}[B]$.

5.1 Special Cases

In this section we apply Theorem 5.5 to two special cases. Namely, we study the departure process, in the large deviations regime, of an M/M/1 queue and a G/D/1 queue.

The departure process of a G/D/1 queue

We assume, as in Section 5, that the interarrival times process $\{A_i, i \in \mathbb{Z}\}$ is stationary and A_i are possibly dependent random variables. The service times B_i are iid random variables

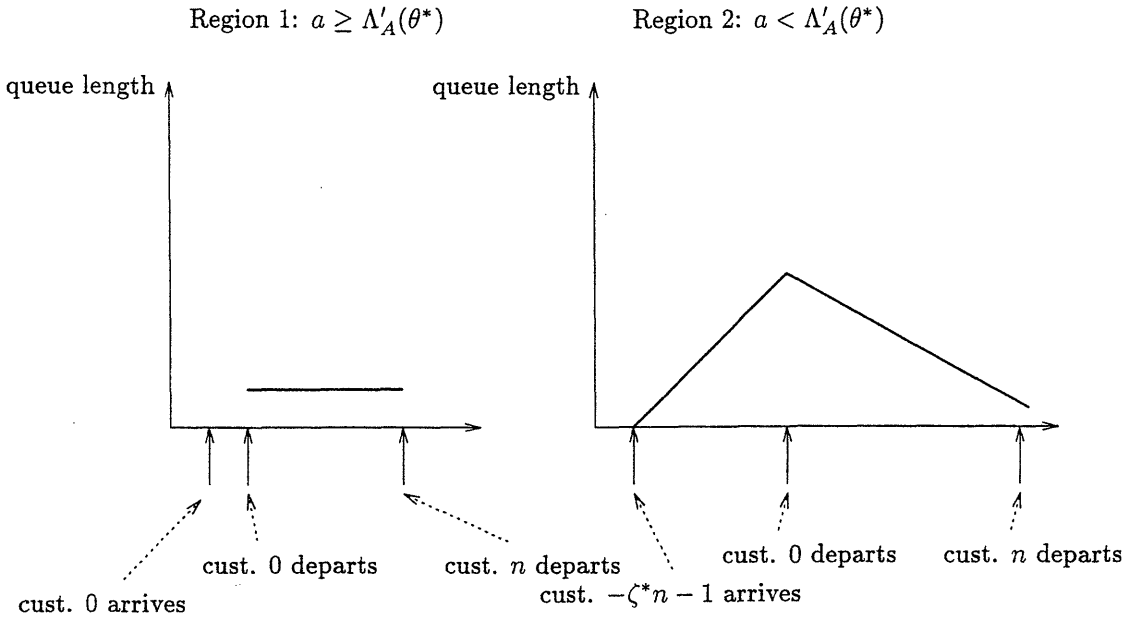

Figure 7: Two cases for the queue length if $a < \mathbf{E}[B]$.

and equal to c w.p.1. Interarrival and service times are assumed independent.

It is straightforward that $\Lambda_B(\theta) = c\theta$. Therefore a simple calculation yields

$$\Lambda_B^{*-}(a) = \begin{cases} +\infty & \text{if } a < c \\ 0 & \text{if } a \geq c \end{cases} \tag{76}$$

Moreover,

$$\Lambda_\Gamma^{*-}(a) = \sup_{\{\theta | \Lambda_A(\theta) - c\theta < 0\}} [\theta a - \Lambda_A^-(\theta)] = \hat{\theta} a - \Lambda_A^-(\hat{\theta}), \tag{77}$$

where $\hat{\theta}$ is the optimizing θ. Note that by taking $a > c$ we have $\Lambda_A(\hat{\theta}) - c\hat{\theta} < 0$, which implies that for such a we have $\Lambda_\Gamma^{*-}(a) = \Lambda_A^{*-}(a)$. Therefore, using Eq. (41),

$$\Lambda_D^{*-}(a) = \begin{cases} +\infty & \text{if } a < c \\ \Lambda_A^{*-}(a) & \text{if } a \geq c \end{cases} \tag{78}$$

This is exactly the result obtained in [dVCW93] for a discrete time model.

The departure process of an M/M/1 queue

We assume that the arrival process is Poisson with rate λ and the service times are iid, distributed according to an exponential distribution with parameter μ.

It is straightforward to calculate

$$\Lambda_A(\theta) = \log\left(\frac{\lambda}{\lambda-\theta}\right), \qquad \Lambda_B(\theta) = \log\left(\frac{\mu}{\mu-\theta}\right). \tag{79}$$

Now, notice that

$$\Lambda_A(\theta) + \Lambda_B(-\theta) = 0 \Leftrightarrow \frac{\lambda}{\lambda-\theta}\frac{\mu}{\mu+\theta} = 1 \Leftrightarrow \theta = 0,\ \theta = \lambda - \mu, \tag{80}$$

which implies that $\theta^* = \lambda - \mu$, where θ^* is defined in the statement of Thm. 4.1. Moreover, notice that

$$\Lambda_A'(\theta^*) = \frac{\lambda-\theta^*}{\lambda}\frac{\lambda}{(\lambda-\theta^*)^2} = \frac{1}{\mu}.$$

Thus, using Eq. (71), we obtain for $a \geq 1/\mu$,

$$\Lambda_D^{*-}(a) = \Lambda_B^{*-}(a) + \Lambda_A^{*-}(a) = \Lambda_A^{*-}(a), \tag{81}$$

since by definition $\Lambda_B^{*-}(a) = 0$ for $a \geq 1/\mu$. Using the second branch of Eq. (71), we obtain for $a < 1/\mu$,

$$\Lambda_D^{*-}(a) = \Lambda_B^{*-}(a) + a(\lambda-\mu) - \log(\lambda/\mu). \tag{82}$$

But

$$\Lambda_B^{*-}(a) = \sup_\theta[\theta a - \Lambda_B^-(\theta)] = a\mu - 1 - \log(a\mu),$$

since, by differentiating, the optimal θ is found equal to $(a\mu-1)/a$. Thus, from Eq. (82), for $a < 1/\mu$,

$$\Lambda_D^{*-}(a) = a\lambda - 1 - \log(a\lambda) = \Lambda_A^{*-}(a). \tag{83}$$

Summarizing Eq. (81) and (83) we finally obtain

$$\Lambda_D^{*-}(a) = \Lambda_A^{*-}(a). \tag{84}$$

This result is in accordance with Burke's output Theorem which states that the departure process of an M/M/1 queue is Poisson with rate λ (see [Kel79]).

6 Superposition of independent streams

In this section we treat the superposition operation of our network model. In particular, we derive a LDP for the process resulting from the superposition of independent arrival streams and we show that the superposition preserves Assumptions B and C. However, as it will become clear in the sequel, in order to derive this LDP we need a result that connects, in

the large deviations regime, the Palm distribution of the arrival process (i.e., as it is seen by a random customer) with its stationary distribution as seen at a random time. This result is presented in Subsection 6.1 and could be of independent interest.

Consider two independent arrival streams. By A_i^1(resp. A_i^2), $i \in \mathbb{Z}$, we denote the interarrival time of the ith customer in stream 1 (resp. 2). We assume that the processes $\{A_i^1, A_i^2,\ i \in \mathbb{Z}\}$ are stationary, and mutually independent. However the interarrival times in each stream may be dependent. We impose Assumptions B and C on the arrival process of each stream. We denote by $A_i^{1,2}$, $i \in \mathbb{Z}$, the interarrival times of the process resulting from the superposition. It should be noted that in order to derive the LDP for the superposition, Assumption C is not used.

The next theorem establishes a LDP for the partial sum $S_{1,n}^{A^{1,2}}$ of the aggregate process, resulting from the superposition of streams 1 and 2.

Theorem 6.1 *Under Assumption B, the partial sum $S_{1,n}^{A^{1,2}}$ of the aggregate process, resulting from the superposition of the independent processes A_i^1, A_i^2, $i \in \mathbb{Z}$, satisfies*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_{1,n}^{A^{1,2}} \leq na] = - \inf_{\substack{\delta_1+\delta_2=1 \\ \delta_1,\delta_2\geq 0}} [\delta_1 \Lambda_{A^1}^{*-}(a/\delta_1) + \delta_2 \Lambda_{A^2}^{*-}(a/\delta_2)] \triangleq -\Lambda_{A^{1,2}}^{*-}(a). \quad (85)$$

Proof : Consult Figure 8. Consider the partial sum $S_{1,n}^{A^{1,2}}$ and let H_1 (resp. H_2) denote

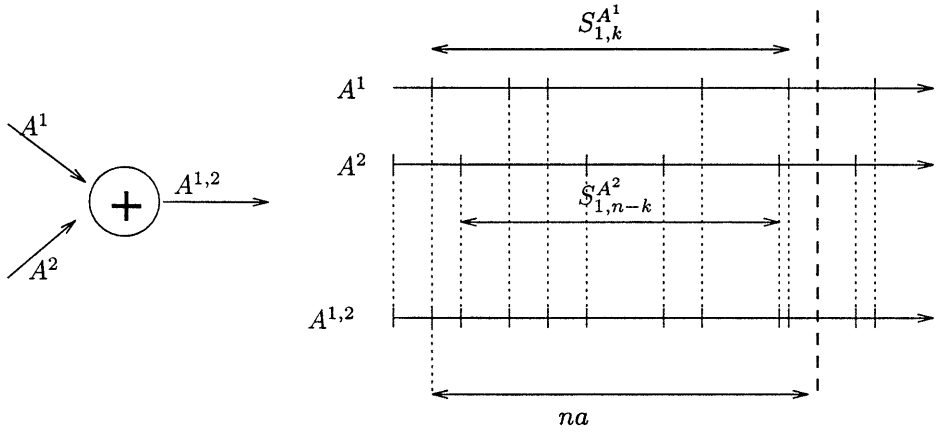

Figure 8: Superposition of two independent streams.

the event that the first customer of the aggregate process originates from stream 1 (resp. 2). We first obtain an upper bound on $\mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_1]$. Notice that

$$\mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_1] = \sum_{k=1}^{n} \mathbf{P}[S_{1,k}^{A^1} \leq na, S_{1,k+1}^{A^1} \geq na] \mathbf{P}_R[S_{1,n-k}^{A^2} \leq na]$$

$$\leq \sum_{k=1}^{n} \mathbf{P}[S_{1,k}^{A^1} \leq na]\mathbf{P}_R[S_{1,n-k}^{A^2} \leq na]. \tag{86}$$

Here, $\mathbf{P}[\cdot]$ denotes the probability distribution seen by a random customer (Palm distribution) and $\mathbf{P}_R[\cdot]$ denotes the probability distribution seen at a random time. Due to the independence of the two arrival streams, an arrival originating from stream 1 constitutes a random incidence in the arrival process of stream 2 and therefore we are interested in the probability distribution seen at a random time for events concerning stream 2.

In Subsection 6.1 it is shown that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}_R[S_{1,n}^{A^2} \leq na] = \lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_{1,n}^{A^2} \leq na] = -\Lambda_{A^2}^{*-}(a). \tag{87}$$

Therefore, from (86), letting $k = n\delta$, $\delta \in [0,1]$ ($n\delta$ is assumed integer), and taking large n we obtain

$$\begin{aligned}
\mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_1] &\leq \sum_{\delta\in[0,1]} \mathbf{P}[S_{1,n\delta}^{A^1} \leq na]\mathbf{P}_R[S_{1,n(1-\delta)}^{A^2} \leq na] \\
&\leq n \sup_{\delta\in[0,1]} \mathbf{P}[S_{1,n\delta}^{A^1} \leq na]\mathbf{P}_R[S_{1,n(1-\delta)}^{A^2} \leq na] \Rightarrow \\
\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_1] &\leq -\inf_{\delta\in[0,1]} [\delta\Lambda_{A^1}^{*-}(a/\delta) + (1-\delta)\Lambda_{A^2}^{*-}(a/(1-\delta))] \\
&= -\inf_{\substack{\delta_1+\delta_2=1 \\ \delta_1,\delta_2\geq 0}} [\delta_1\Lambda_{A^1}^{*-}(a/\delta_1) + \delta_2\Lambda_{A^2}^{*-}(a/\delta_2)]. \qquad (88)
\end{aligned}$$

To obtain a lower bound notice that

$$\begin{aligned}
\mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_1] &\geq \sup_{\delta\in[0,1]} \mathbf{P}[S_{1,n\delta}^{A^1} \leq na]\mathbf{P}_R[S_{1,n(1-\delta)}^{A^2} \leq na] \Rightarrow \\
\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_1] &\geq -\inf_{\delta\in[0,1]} [\delta\Lambda_{A^1}^{*-}(a/\delta) + (1-\delta)\Lambda_{A^2}^{*-}(a/(1-\delta))] \\
&= -\inf_{\substack{\delta_1+\delta_2=1 \\ \delta_1,\delta_2\geq 0}} [\delta_1\Lambda_{A^1}^{*-}(a/\delta_1) + \delta_2\Lambda_{A^2}^{*-}(a/\delta_2)]. \qquad (89)
\end{aligned}$$

Finally, observe that because of symmetry, Eqs. (88) and (89) also hold for $\mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_2]$. This along with the fact

$$\mathbf{P}[S_{1,n}^{A^{1,2}} \leq na] = \mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_1]\mathbf{P}[H_1] + \mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_2]\mathbf{P}[H_2],$$

proves the theorem.

■

Remark : Let δ_1^*, δ_2^* solve the optimization problem in (85). It can be seen that the most likely path to have a large deviation in the aggregate process is to maintain an empirical arrival rate of $\frac{\delta_1^*}{a}$ in stream 1 and a rate of $\frac{\delta_2^*}{a}$ in stream 2. Then, since $\delta_1^* + \delta_2^* = 1$ the

empirical rate of the aggregate process is $\frac{1}{a}$.

Using induction on the number of streams superimposed we generalize Theorem 6.1 to obtain the following corollary.

Corollary 6.2 *Under Assumption B, the partial sum $S_{1,n}^{A^1,\ldots,m}$ of the aggregate process, resulting from the superposition of the m independent processes $A_i^1,\ldots,A_i^m$ $i\in\mathbb{Z}$, satisfies*

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbf{P}[S_{1,n}^{A^1,\ldots,m}\leq na]=-\inf_{\substack{\delta_1+\cdots+\delta_m=1\\ \delta_1,\ldots,\delta_m\geq 0}}\sum_{k=1}^{m}\delta_k\Lambda_{A^k}^{*-}(a/\delta_k)\stackrel{\triangle}{=}-\Lambda_{A^1,\ldots,m}^{*-}(a). \tag{90}$$

Using convex duality one can obtain the limiting log-moment generating function $\Lambda_{A^1,\ldots,m}^{-}(\cdot)$ of $S_{1,n}^{A^1,\ldots,m}$ as the convex dual of its large deviations rate function $\Lambda_{A^1,\ldots,m}^{*-}(\cdot)$.

We now proceed into proving that the aggregate process, resulting from the superposition of independent streams which satisfy Assumptions B and C also satisfies the same assumptions. It is clear that the process resulting from the superposition satisfies Assumption B, since we have proven a LDP for this process with large deviations rate function expressed as a function of the large deviations rate function of the superimposed processes. The next Theorem establishes that the process resulting from the superposition satisfies Assumption C.

Theorem 6.3 *Assume that the m independent processes $A_i^1,\ldots,A_i^m$ $i\in\mathbb{Z}$ satisfy Assumption C. The aggregate process resulting from their superposition also satisfies Assumption C.*

Proof : It suffices to prove the result for $m=2$ since by using induction we can prove it for any m. We need to prove that for every $\epsilon_1,\epsilon_2,a>0$, there exists M_S such that for all $n\geq M_S$

$$e^{-n(\Lambda_{A^{1,2}}^{*-}(a)+\epsilon_2)}\leq\mathbf{P}[S_{1,j}^{A^{1,2}}-ja\leq\epsilon_1 n,\ j=1,\ldots,n]. \tag{91}$$

Following the steps of the proof of Theorem 6.1 we consider the scenario that a fraction δ of customers of the aggregate process originates from the A^1 process. Again, H_1 denotes the event that customer 1 of the aggregate process originates from the A^1 process. We have

$$\begin{aligned}\mathbf{P}[S_{1,j}^{A^{1,2}}-ja\leq\epsilon_1 n,\ j=1,\ldots,n\mid H_1]\geq\\ \geq\sup_{\delta\in[0,1]}\mathbf{P}[S_{1,j\delta}^{A^1}-ja\leq\epsilon_1 n,\ j=1,\ldots,n]\times\\ \mathbf{P}_R[S_{1,j(1-\delta)}^{A^2}-ja\leq\epsilon_1 n,\ j=1,\ldots,n].\end{aligned} \tag{92}$$

Using Assumption C for the A^1 stream we obtain for large enough n

$$\mathbf{P}[S_{1,j\delta}^{A^1}-ja\leq\epsilon_1 n,\ j=1,\ldots,n]\geq e^{-n\delta(\Lambda_{A^1}^{*-}(a/\delta)+\epsilon')}. \tag{93}$$

In Subsection 6.1 (Lemma 6.6) it is shown that for large enough n

$$\mathbf{P}_R[S^{A^2}_{1,j(1-\delta)} - ja \leq \epsilon_1 n,\ j = 1, \ldots, n] \geq e^{-n(1-\delta)(\Lambda^{*-}_{A^2}(a/(1-\delta))+\epsilon'')}. \tag{94}$$

To obtain (91) it suffices to choose appropriate ϵ' and ϵ'' such that for large enough n and given ϵ_2

$$e^{-n\inf_{\delta\in[0,1]}[\delta(\Lambda^{*-}_{A^1}(a/\delta)+\epsilon')+(1-\delta)(\Lambda^{*-}_{A^2}(a/(1-\delta))+\epsilon'')]} \geq e^{-n(\Lambda^{*-}_{A^{1,2}}(a)+\epsilon_2)}.$$

■

6.1 Connection between Palm and stationary distributions in the large deviations regime

In this subsection we show that the stationary and the Palm distribution of the same point process have the same large deviations behaviour.

Consider a stationary arrival process satisfying Assumptions B with interarrivals A_i, $i \in \mathbb{Z}$. Due to Assumption B and the Gärtner-Ellis Thm. we have

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S^A_{1,n} \leq na] = -\Lambda^{*-}_A(a). \tag{95}$$

As explained in the proof of Thm. 6.1, $\mathbf{P}[\cdot]$ denotes the probability distribution seen by a random customer (customer 1 in the case of Eq. (95)). Consider now a random time (say $t = 0$) and assume that customer 1 is the first customer to arrive after $t = 0$. Let U, V denote the duration and the age, respectively, of A_0. The situation is depicted in Figure 9. By $\mathbf{P}_R[\cdot]$ we denote the probability distribution seen at the random time $t = 0$ and we

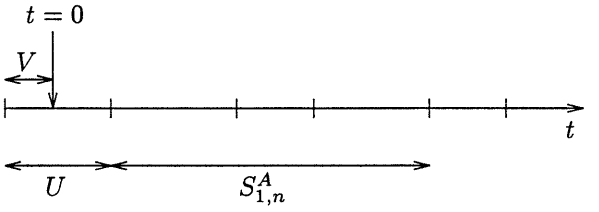

Figure 9: The arrival process seen at a random time.

are interested in obtaining a LDP for $S^A_{1,n}$ under $\mathbf{P}_R[\cdot]$. The next theorem establishes the result. Moreover, we are also interested in obtaining a LDP result for the partial sum process $\{S^A_{1,j},\ j = 1, \ldots, n\}$ under $\mathbf{P}_R[\cdot]$ when Assumption C is satisfied. The latter result is obtained in Lemma 6.6.

Theorem 6.4 *Under Assumption B we have*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}_R[S^A_{1,n} \leq na] = -\Lambda^{*-}_A(a). \tag{96}$$

Proof : Let $\mathbf{E}_R[\cdot]$ denote the expectation with respect to $\mathbf{P}_R[\cdot]$. We use a standard procedure to relate $\mathbf{E}_R[\cdot]$ to $\mathbf{E}[\cdot]$ (see [Wal88]). Consider an arbitrary function $f(\cdot)$ of $S^A_{1,n}$. It can be shown ([Wal88, ch. 7]) that

$$\mathbf{E}_R[f(S^A_{1,n}) \mid V = v, U = u] = \mathbf{E}[f(S^A_{1,n}) \mid A_0 = u].$$

Thus following the steps in [Wal88, ch. 7],

$$\begin{aligned} \mathbf{E}_R[f(S^A_{1,n})] &= \frac{1}{\mathbf{E}[A_1]} \int_{u=0}^{\infty} \int_{v=0}^{u} \mathbf{E}[f(S^A_{1,n}) \mid A_0 = u]\, dv\, dF_{A_0}(u) \\ &= \mathbf{E}[\int_{v=0}^{A_0} f(S^A_{1,n})\, dv] \\ &= \mathbf{E}[A_0 f(S^A_{1,n})], \end{aligned} \tag{97}$$

where we have assumed without loss of generality that $\mathbf{E}[A_1] = 1$, and we have used the notation $F_{A_0}(\cdot)$ for the distribution function of A_0.

To obtain an upper bound on $\mathbf{E}_R[e^{\theta S^A_{1,n}}]$ we set $f(\cdot) = e^{\theta \cdot}$ and use Hölder's inequality. Namely,

$$\begin{aligned} & \mathbf{E}_R[e^{\theta S^A_{1,n}}] = \mathbf{E}[A_0 e^{\theta S^A_{1,n}}] \\ (p+q=1) \qquad & = \mathbf{E}[(A_0^p)^{1/p} (e^{\theta q S^A_{1,n}})^{1/q}] \leq \mathbf{E}[A_0^p]^{1/p} \mathbf{E}[e^{\theta q S^A_{1,n}}]^{1/q}, \end{aligned} \tag{98}$$

which implies

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}_R[e^{\theta S^A_{1,n}}] \leq \limsup_{n\to\infty} \frac{\log \mathbf{E}[A_0^p]}{pn} + \frac{\Lambda_A(\theta q)}{q} = \frac{\Lambda_A(\theta q)}{q}, \tag{99}$$

since the first term of the right hand side vanishes for $p \neq 0$. Taking the limit as $q \to 0$ in the above equation, using Hospital's rule and the convexity of $\Lambda_A(\cdot)$ we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}_R[e^{\theta S^A_{1,n}}] \leq \limsup_{q\to 0} \frac{\Lambda_A(\theta q)}{q} = \theta \Lambda'_A(0) \leq \Lambda_A(\theta). \tag{100}$$

Therefore using Eq. (100) and the Markov inequality we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}_R[S^A_{1,n} \leq na] \leq -\Lambda^{*-}_A(a). \tag{101}$$

To obtain now a lower bound on $\mathbf{P}_R[S^A_{1,n} \leq na]$ set $f(S^A_{1,n}) = \mathbf{1}\{S^A_{1,n} \leq na\}$ in Eq. (97),

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. We have

$$\begin{aligned}\mathbf{P}_R[S^A_{1,n} \le na] &= \int_0^\infty u\mathbf{P}[S^A_{1,n} \le na \mid A_0 = u]\, dF_{A_0}(u) \\ &\ge \frac{1}{n^2}\int_{1/n^2}^\infty \mathbf{P}[S^A_{1,n} \le na \mid A_0 = u]\, dF_{A_0}(u) \\ &= \frac{1}{n^2}\mathbf{P}[S^A_{1,n} \le na, A_0 \ge \tfrac{1}{n^2}]. \end{aligned} \tag{102}$$

We need the following lemma the proof of which is deferred until the end of the current proof.

Lemma 6.5 *Under Assumption B and for every positive ϵ and a, there exists $N_{a,\epsilon}$ such that for every $n \ge N_{a,\epsilon}$ it holds*

$$\mathbf{P}[S^A_{1,n} \le na, A_0 \ge \tfrac{1}{n^2}] \ge e^{-n(\Lambda^{*-}_A(a)+\epsilon)}. \tag{103}$$

We now use Lemma 6.5 in Eq. (102) and take $\epsilon \to 0$ to obtain

$$\liminf_{n\to\infty} \frac{1}{n}\log \mathbf{P}_R[S^A_{1,n} \le na] \ge -\Lambda^{*-}_A(a).$$

■

Proof of Lemma 6.5: Eq. (95) implies that for every positive ϵ' and a there exists $N'_{a,\epsilon'}$ such that for every $n \ge N'_{a,\epsilon'}$ it holds

$$e^{-n(\Lambda^{*-}_A(a)+\epsilon')} \le \mathbf{P}[S^A_{1,n} \le na] \le e^{-n(\Lambda^{*-}_A(a)-\epsilon')}. \tag{104}$$

Fix now $a, \epsilon' > 0$, and let $\delta = \epsilon'$. We have

$$\begin{aligned} &\mathbf{P}[S^A_{1,n} \le na, A_0 \ge \tfrac{1}{n^2}] = \\ \text{(by stationarity)} \quad &= \frac{1}{n\delta}\sum_{i=1}^{n\delta}\mathbf{P}[S^A_{i+1,i+n} \le na, A_i \ge \tfrac{1}{n^2}] \\ \text{(union bound)} \quad &\ge \frac{1}{n\delta}\mathbf{P}[\exists i \in [1,n\delta] \text{ s.t. } S^A_{i+1,i+n} \le na, A_i \ge \tfrac{1}{n^2}] \\ &\ge \frac{1}{n\delta}\mathbf{P}[S^A_{1,n(1+\delta)} \le na, \exists i \in [1,n\delta] \text{ s.t. } A_i \ge \tfrac{1}{n^2}] \\ &\ge \frac{1}{n\delta}\mathbf{P}[S^A_{1,n(1+\delta)} \le na, \sum_{i=1}^{n\delta} A_i \ge \tfrac{n\delta}{n^2}] \\ (\mathbf{P}[A\cap B] \ge \mathbf{P}[A] - \mathbf{P}[B^C]) \quad &\ge \frac{1}{n\delta}\mathbf{P}[S^A_{1,n(1+\delta)} \le na] - \frac{1}{n\delta}\mathbf{P}[S^A_{1,n\delta} \le \tfrac{\delta}{n}] \\ &\ge \frac{1}{n\delta}e^{-n(1+\delta)(\Lambda^{*-}_A(\frac{a}{1+\delta})+\epsilon')} - \frac{1}{n\delta}\mathbf{P}[S^A_{1,n\delta} \le \tfrac{\delta}{n}], \end{aligned} \tag{105}$$

where the last inequality holds for all $n \geq N'_{\frac{a}{1+\delta},\epsilon'}$. Note that we have used the notation B^C to denote the complement of B. We next show that for $n \to \infty$ (keeping a, δ, ϵ' fixed) we can neglect the second term in the right hand side of (105). To see that note that for all β positive there exists $N_{\beta,\epsilon'}$ such that for all $n \geq N_{\beta,\epsilon'}$ it holds

$$\mathbf{P}[S^A_{1,n\delta} \leq \tfrac{\delta}{n}] \leq \mathbf{P}[S^A_{1,n\delta} \leq n\delta\beta] \leq e^{-n(\Lambda^{*-}_A(\beta)-\epsilon')}. \tag{106}$$

By taking β, δ and ϵ' small enough and $n \geq N_{\beta,\epsilon'}$ we can achieve

$$\Lambda^{*-}_A(\beta) - \epsilon' > (1+\delta)(\Lambda^{*-}_A(\tfrac{a}{1+\delta}) + \epsilon'). \tag{107}$$

Here we are using the fact that for sufficiently small β, $\Lambda^{*-}_A(\beta) > \Lambda^{*-}_A(\frac{a}{1+\delta})$ since $\Lambda^{*-}_A(\beta)$ is monotonically increasing as $\beta \downarrow 0$.

Observe now that the value of β which satisfies Eq. (107) is a function of a, δ and ϵ'. Therefore, using Eq. (106), there exists $N_{a,\delta,\epsilon'}$ such that for all $n \geq N_{a,\delta,\epsilon'}$ it holds

$$-\frac{1}{2n\delta}\mathbf{P}[S_{1,n\delta} \leq \tfrac{\delta}{n}] \geq -\frac{1}{2n\delta}e^{-n(1+\delta)(\Lambda^{*-}_A(\frac{a}{1+\delta})+\epsilon')}. \tag{108}$$

Combining Eqs. (108) and (105) we conclude that there exists $\hat{N}_{a,\delta,\epsilon'}$ such that for all $n \geq \hat{N}_{a,\delta,\epsilon'}$ it holds

$$\mathbf{P}[S^A_{1,n} \leq na, A_0 \geq \tfrac{1}{n^2}] \geq \frac{1}{2n\delta}e^{-n(1+\delta)(\Lambda^{*-}_A(\frac{a}{1+\delta})+\epsilon')}. \tag{109}$$

We now choose ϵ' such that (recall $\delta = \epsilon'$)

$$\frac{1}{2n\epsilon'}e^{-n(1+\epsilon')(\Lambda^{*-}_A(\frac{a}{1+\epsilon'})+\epsilon')} \geq e^{-n(\Lambda^{*-}_A(a)+\epsilon)},$$

for all $n \geq N_{a,\epsilon}$. This can be done due to the continuity of $\Lambda^{*-}_A(\cdot)$.

■

Lemma 6.6 *Under Assumptions B and C we have that for every $\epsilon_1, \epsilon_2, a > 0$ there exists $N_{a,\epsilon_1,\epsilon_2}$ such that for all $n \geq N_{a,\epsilon_1,\epsilon_2}$ it holds*

$$\mathbf{P}_R[S^A_{1,j} \leq ja + \epsilon_1 n,\ j = 1, \ldots, n] \geq e^{-n(\Lambda^{*-}_A(a)+\epsilon_2)}. \tag{110}$$

Proof : Following the proof of the lower bound in Thm. 6.4 (Eq. (102)) we have

$$\mathbf{P}_R[S^A_{1,j} \leq ja + \epsilon_1 n,\ j = 1, \ldots, n] \geq \frac{1}{n^2}\mathbf{P}[S^A_{1,j} \leq ja + \epsilon_1 n,\ j = 1, \ldots, n,\ A_0 \geq \tfrac{1}{n^2}]. \tag{111}$$

Now, as in the proof of Lemma 6.5, fixing $a, \epsilon_1, \epsilon_2 > 0$ we obtain

$$\begin{aligned}
&\mathbf{P}[S^A_{1,j} \leq ja + \epsilon_1 n,\ j = 1, \dots, n,\ A_0 \geq \tfrac{1}{n^2}] = \\
&= \frac{1}{n\delta} \sum_{k=1}^{n\delta} \mathbf{P}[S^A_{1+k,j+k} \leq ja + \epsilon_1 n,\ j = 1, \dots, n,\ A_k \geq \tfrac{1}{n^2}] \\
&\geq \frac{1}{n\delta} \mathbf{P}[\exists k \in [1, n\delta] \text{ s.t. } S^A_{1+k,j+k} \leq ja + \epsilon_1 n,\ j = 1, \dots, n,\ A_k \geq \tfrac{1}{n^2}] \\
&\geq \frac{1}{n\delta} \mathbf{P}[\forall k \in [1, n\delta]\ S^A_{1+k,j+k} \leq ja + \epsilon_1 n,\ j = 1, \dots, n, \exists k \in [1, n\delta] \text{ s.t. } A_k \geq \tfrac{1}{n^2}] \\
&\geq \frac{1}{n\delta} \mathbf{P}[\forall k \in [1, n\delta]\ S^A_{1+k,j+k} \leq ja + \epsilon_1 n,\ j = 1, \dots, n] - \frac{1}{n\delta} \mathbf{P}[S^A_{1,n\delta} \leq \tfrac{\delta}{n}]. \qquad (112)
\end{aligned}$$

Now notice that

$$\begin{aligned}
&\mathbf{P}[\forall k \in [1, n\delta]\ S^A_{1+k,j+k} \leq ja + \epsilon_1 n,\ j = 1, \dots, n] = \\
&\quad \mathbf{P}[\forall k \in [1, n\delta]\ S^A_{1,j+k} - S^A_{1,k} \leq (j+k)a - ka + \epsilon_1 n,\ j = 1, \dots, n] \geq \\
&\quad \mathbf{P}[S^A_{1,j+k} \leq (j+k)a + \tfrac{\epsilon_1 n}{2},\ \forall k \in [1, n\delta],\ j = 1, \dots, n,\ S^A_{1,k} \geq ka - \tfrac{\epsilon_1 n}{2},\ \forall k \in [1, n\delta]] = \\
&\quad \mathbf{P}[S^A_{1,j+k} \leq (j+k)a + \tfrac{\epsilon_1 n}{2},\ \forall k \in [1, n\delta],\ j = 1, \dots, n] \geq e^{-n(1+\delta)(\Lambda^{*-}_A(a) + \epsilon')}, \qquad (113)
\end{aligned}$$

where the last equality is obtained by choosing sufficiently small δ such that $n\delta a - \frac{\epsilon_1 n}{2} < 0$ which implies that $\mathbf{P}[S^A_{1,k} \geq ka - \frac{\epsilon_1 n}{2},\ \forall k \in [1, n\delta]] = 1$. The last inequality holds, due to Assumption C, for all $n \geq N'_{a,\epsilon_1,\epsilon'}$. Now, as in Lemma 6.5 it can be shown that there exists $N''_{a,\delta,\epsilon'}$ such that for all $n \geq N''_{a,\delta,\epsilon'}$ it holds

$$-\frac{1}{2n\delta} \mathbf{P}[S_{1,n\delta} \leq \tfrac{\delta}{n}] \geq -\frac{1}{2n\delta} e^{-n(1+\delta)(\Lambda^{*-}_A(a) + \epsilon')}. \qquad (114)$$

Combining Eqs. (111), (112), (113) and (114) we conclude that there exists $\hat{N}_{a,\epsilon_1,\delta,\epsilon'}$ such that for all $n \geq \hat{N}_{a,\epsilon_1,\delta,\epsilon'}$ it holds

$$\mathbf{P}_R[S^A_{1,j} \leq ja + \epsilon_1 n,\ j = 1, \dots, n] \geq \frac{1}{2n^3\delta} e^{-n(1+\delta)(\Lambda^{*-}_A(a) + \epsilon')}. \qquad (115)$$

We now choose ϵ' and if necessary δ smaller than the one chosen above for the purposes of (113), such that

$$\frac{1}{2n^3\delta} e^{-n(1+\delta)(\Lambda^{*-}_A(a) + \epsilon')} \geq e^{-n(\Lambda^{*-}_A(a) + \epsilon_2)},$$

for $n \geq N_{a,\epsilon_1,\epsilon_2}$.

■

7 Bernoulli splitting of a stream

In this section we treat the splitting operation of our network model. In particular, we derive a LDP for the process resulting from the splitting of a stream to a number of streams and we show that splitting preserves Assumptions B and C.

Consider a stream with stationary interarrival times A_i, $i \in \mathbb{Z}$, which is split to 2 substreams. In particular, arrivals of the *"master"* stream are directed with probabilities p and $1-p$ to substreams 1 and 2, respectively. The next theorem provides a LDP for stream 1. Since stream 1 is chosen arbitrarily, by relabelling the streams one can obtain a LDP for stream 2. The more general case in which the master stream is split to more than two substreams can be handled successive splitting to two substreams. Let us denote by A_i^1, A_i^2, $i \in \mathbb{Z}$, the interarrival times of substreams $1, 2$, respectively. $\Lambda_A^*(\cdot)$ and $\Lambda_A(\cdot)$ denote the large deviations rate function and the limiting log-moment generating function of the master stream.

Theorem 7.1 *Under Assumption B, the partial sum $S_{1,n}^{A^1}$ of substream 1 satisfies*

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log \mathbf{P}[S_{1,n}^{A^1} \le na] &= -\inf_{\delta\ge 0}\left[(1+\delta)\Lambda_A^{*-}\left(\frac{a}{1+\delta}\right) - \right. \\ &\left. -\log(1+\delta) - \log p - \delta\log(1-p)\right] \triangleq -\Lambda_{A^1}^{*-}(a). \quad (116)\end{aligned}$$

Proof : Let out of $n+k$ arrivals from the master stream, n be directed to substream 1 and k to substream 2, respectively. Thus,

$$\begin{aligned}&\mathbf{P}[S_{1,n}^{A^1} \le na] = \\ &= \sum_{k=0}^{\infty}\binom{n-1+k}{n-1}\mathbf{P}[S_{1,n+k}^{A} \le na]p^n(1-p)^k \\ &= \sum_{\delta\ge 0}\binom{n(1+\delta)-1}{n-1}\mathbf{P}[S_{1,n(1+\delta)}^{A} \le na]p^n(1-p)^{n\delta}, \quad (117)\end{aligned}$$

where we have made the substitution $k = n\delta$ (assuming $n\delta$ is integer). Using Stirling's approximation we obtain

$$\binom{n(1+\delta)-1}{n-1} \sim (n^{-1/2}(1+\delta)^n).$$

Thus, using (117) we obtain

$$\mathbf{P}[S_{1,n}^{A_1} \le na] \sim \sup_{\delta\ge 0} n^{-1/2}(1+\delta)^n\mathbf{P}[S_{1,n(1+\delta)}^{A} \le na]p^n(1-p)^{n\delta}. \quad (118)$$

Finally, by using

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbf{P}[S^A_{1,n(1+\delta)}\leq na]=-(1+\delta)\Lambda^{*-}_A\left(\frac{a}{1+\delta}\right),$$

we obtain (116).

■

We now argue that splitting preserves Assumptions B, and C. It is clear that the process resulting from splitting satisfies Assumption B, since we have proven a LDP for this process with large deviations rate function expressed as a function of the large deviations rate function of the master process. The next theorem establishes that the process resulting from splitting satisfies Assumption C.

Theorem 7.2 *Assume that the process* $\{A_i,\ i\in\mathbb{Z}\}$, *satisfies Assumptions B and C. Then the* A^1 *process satisfies Assumption C.*

Proof : The proof is very similar to the proof of Theorem 7.1. As in Eq. (117) we consider the scenario that a fraction $\frac{1}{1+\delta}$ of the arrivals from the master stream is routed to stream 1. Thus, to achieve $S^{A^1}_{1,j}-ja\leq\epsilon_1 n,\ j=1,\ldots,n$, it suffices to have $S^A_{1,j(1+\delta)}-ja\leq\epsilon_1 n,\ j=1,\ldots,n$. Taking also into account the number of ways that the routing of n arrivals to stream 1 from a total of $n(1+\delta)$ from the master stream can be done along with the corresponding routing probabilities we obtain

$$\begin{aligned}&\mathbf{P}[S^{A^1}_{1,j}-ja\leq\epsilon_1 n,\ j=1,\ldots,n]\geq\\&\quad\geq\sup_{\delta\geq 0}(1+\delta)^n\mathbf{P}[S^A_{1,j(1+\delta)}-ja\leq\epsilon_1 n,\ j=1,\ldots,n]p^n(1-p)^{n\delta},\end{aligned}$$

and proceed as in Theorem 7.1, using Assumption C for the master process.

■

8 An Example: Queues in Tandem

In this section we apply the results derived so far to obtain LDP's for two G/GI/1 queues in tandem. Moreover we work out a numerical example in order to get a qualitative understanding of the results. Large deviations results for tandem queues with renewal arrivals and exponential servers have been reported in [GA94].

Consider two G/GI/1 queues in tandem. Let $A_i,\ i\in\mathbb{Z}$, denote the interarrival times in the first queue and $B^1_i, B^2_i,\ i\in\mathbb{Z}$, the service times in the first and second queue respectively. These processes are mutually independent, stationary and satisfy Assumptions B and C.

According to Corollary 5.6 the limiting log-moment generating function of the departure

process from the first queue is given by

$$\Lambda_D^-(\theta) = \begin{cases} \inf_{x+y=\theta}\{\Lambda_{B^1}^-(x) + \Lambda_A^-(y)\} & \text{if } \theta \geq \hat{\theta} \\ \Lambda_{B^1}^-(\theta - \theta^*) + \Lambda_A(\theta^*) & \text{if } \theta < \hat{\theta} \end{cases} \tag{119}$$

where

$$\hat{\theta} \triangleq \frac{d}{da}\left[\Lambda_{B^1}^{*-}(a) + \Lambda_A^{*-}(a)\right]_{a=\Lambda_A'(\theta_1^*)}.$$

Applying Theorem 4.1 we obtain that the tail probability of the stationary waiting time, W_2, seen by a customer in the second queue, is characterized by

$$\mathbf{P}[W_2 \geq U] \sim e^{\theta_2^* U}, \tag{120}$$

where U is large enough and $\theta_2^* < 0$ is the smallest root of the equation $\Lambda_D(\theta) + \Lambda_{B^2}(-\theta) = 0$. Since for $\theta \leq 0$ the equation $\Lambda_D(\theta) + \Lambda_{B^2}(-\theta) = 0$ has exactly the same roots as the equation $\Lambda_D^-(\theta) + \Lambda_{B^2}^+(-\theta) = 0$, it turns out that θ_2^* is the smallest root of the equation

$$\begin{array}{ll} \inf_{x+y=\theta}\{\Lambda_{B^1}^-(x) + \Lambda_A^-(y)\} + \Lambda_{B^2}^+(-\theta) = 0 & \text{if } \theta \geq \hat{\theta} \\ \Lambda_{B^1}^-(\theta - \theta^*) + \Lambda_A(\theta^*) + \Lambda_{B^2}^+(-\theta) = 0 & \text{if } \theta < \hat{\theta} \end{array}$$

It is instructive to characterize the most likely path along which the LDP for the waiting time occurs in the second queue. The remarks after the proof of Theorem 4.1, suggest that the most likely path for the waiting time in the second queue is characterized by

$$\mathbf{P}[W_0^2 \geq (i+1)a] \sim \sup_{x_2 - x_1 = a} \mathbf{P}[S_{-i,0}^D \leq (i+1)x_1]\mathbf{P}[S_{-i-1,-1}^{B^2} \geq (i+1)x_2], \tag{121}$$

where W_0^2 denotes the waiting time of the 0th customer in the second queue and i is large enough. Setting $U = (i+1)a$, we obtain for large enough U

$$\mathbf{P}[W_0^2 \geq U] \sim \exp\left\{-U \inf_a \frac{1}{a} \inf_{x_2 - x_1 = a}[\Lambda_D^{*-}(x_1) + \Lambda_{B^2}^{*+}(x_2)]\right\}. \tag{122}$$

Let (a^*, x_1^*, x_2^*) be the optimal solution of the optimization problem appearing in (122). Eq. (122) suggests that the waiting time in the second queue builds up by maintaining an empirical rate of $1/x_1^*$ for the process D (departure from first queue) and an empirical service rate (process B^2) of $1/x_2^*$.

We use the remarks after Theorem 5.5 to characterize the most likely path for the process D to maintain an empirical rate of $1/x_1^*$. Let i^* be defined by the equation $i^* + 1 = U/a^*$. From (121), it can be seen that it suffices to characterize the most likely path along which the event $\{S_{-i^*,0}^D \leq (i^*+1)x_1^*\}$ occurs. As shown in Theorem 5.5, this most likely path is

characterized by

$$\mathbf{P}[S^D_{-i^*,0} \leq (i^*+1)x_1^*] \sim \exp\left\{(i^*+1)\sup_{\zeta\geq 0}\sup_{y_1-y_2=a}\left[-(1+\zeta)\Lambda_A^{*-}\left(\frac{y_1}{1+\zeta}\right)-\right.\right.$$
$$\left.\left.-\zeta\Lambda_{B^1}^{*+}\left(\frac{y_2}{\zeta}\right)\right]-(i^*+1)\Lambda_{B^1}^{*-}(a)\right\}. \quad (123)$$

Let (y_1^*, y_2^*, ζ^*) be the solution of the optimization problem appearing in Eq. (123). We depict the most likely path in Figure 10, for the case $x_1^* < \mathbf{E}[B^1]$, where $\mathbf{E}[B^1]$ denotes the expected service time in the first queue. For the case $x_1^* \geq \mathbf{E}[B^1]$ the most likely path can be similarly identified (see the discussion after the proof of Lemma 5.4).

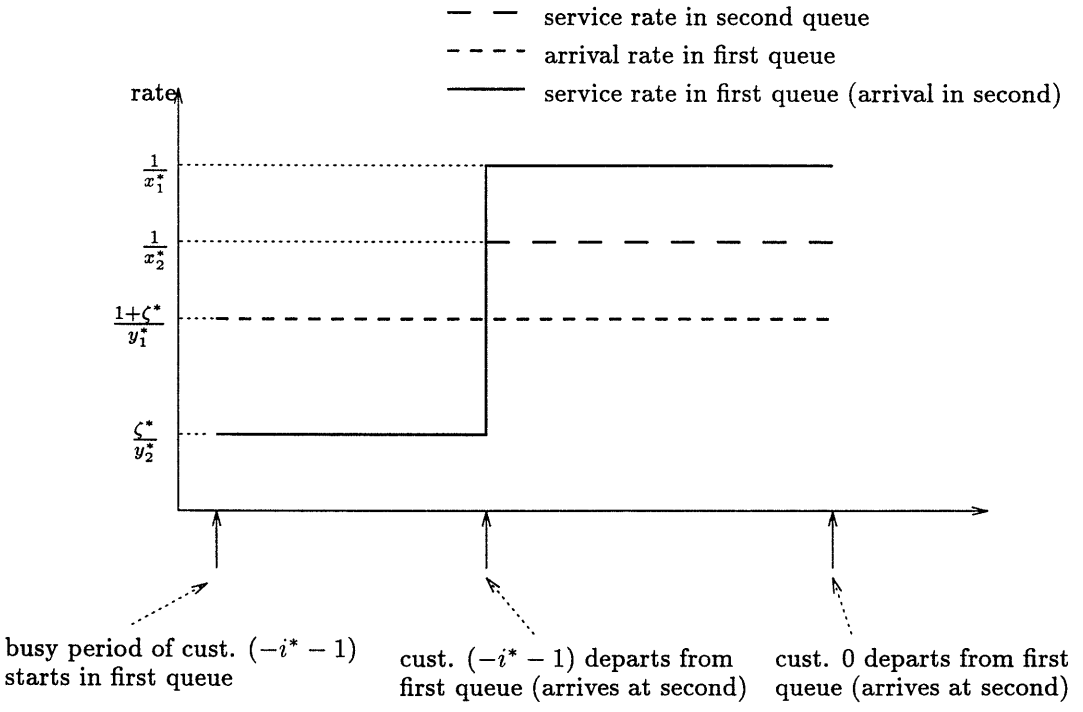

Figure 10: The most likely path for the waiting time in the second queue.

We now proceed with a numerical example. We chose the arrival process A to be a two-state *Markov modulated* deterministic process. More precisely, we consider a two-state Markov chain with transition probability matrix

$$P = \begin{bmatrix} 1/3 & 2/3 \\ 1/2 & 1/2 \end{bmatrix}, \quad (124)$$

and we let the interarrival times be equal to $\frac{1}{\lambda_1} = \frac{1}{5}$ w.p.1 when the chain is at state 1, and equal to $\frac{1}{\lambda_2} = \frac{1}{2}$ w.p.1 when the chain is at state 2. The steady-state probability vector for this Markov chain is $[\pi_1 \ \pi_2] = [\frac{3}{7} \ \frac{4}{7}]$ and thus the mean arrival rate is $\lambda_1\pi_1 + \lambda_2\pi_2 = 3.29$.

We chose a deterministic server for both queues 1 and 2 with rate $c = 3.87$.

Theorem 3.1.2 in [DZ93b] calculates the limiting log-moment generating function for the arrival process as the largest eigenvalue of the matrix $P_\theta \triangleq [p_{ij}e^{\theta/\lambda_j}]$ which in our case is

$$P = \begin{bmatrix} 1/3e^{\theta/5} & 2/3e^{\theta/2} \\ 1/2e^{\theta/5} & 1/2e^{\theta/2} \end{bmatrix}. \tag{125}$$

We performed several calculations using the software package *Matlab*. For the tail probability of the waiting time in the first queue we found that $\theta_1^* = -18.62$. We calculated the large deviations rate functions $\Lambda_A^{*-}(a)$ and $\Lambda_D^{*-}(a)$ for the arrival process and the departure process from the first queue, respectively. The results appear in Figure 11. To calculate $\Lambda_D^{*-}(a)$ we used Eq. (78). It can be seen that the first queue has a smoothing effect on the

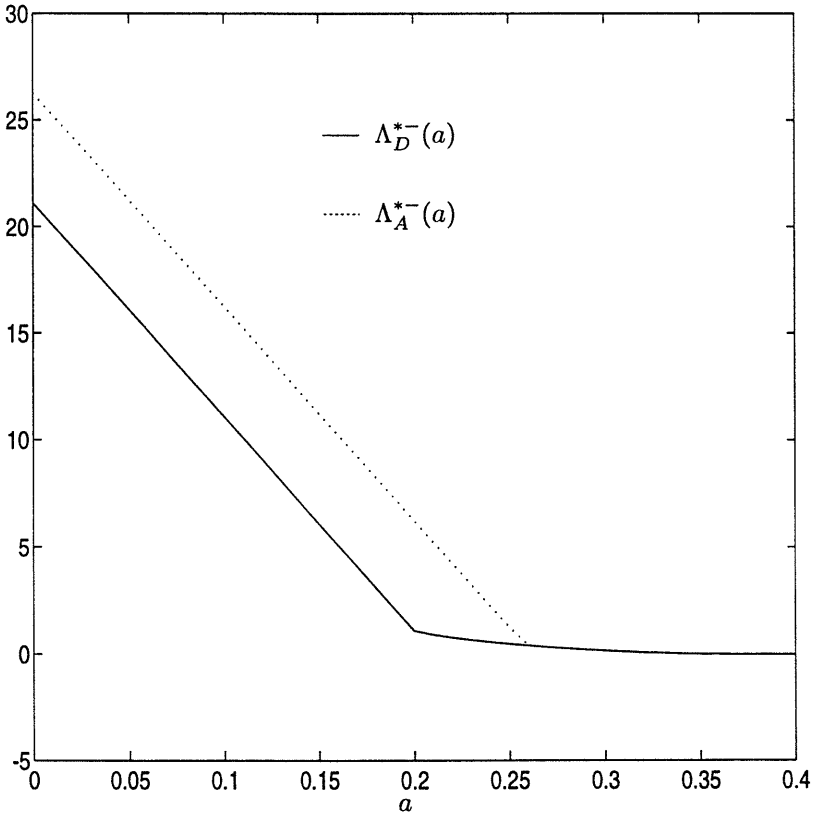

Figure 11: $\Lambda_A^{*-}(a)$ and $\Lambda_D^{*-}(a)$ for the numerical example.

arrival process. In other words, the departure process deviates from its mean with smaller probability than the arrival process does. We also found that $\Lambda_D(\theta) + \Lambda_{B^2}(-\theta)$ is strictly negative for all $\theta < 0$, so as it can be seen from the proof of Theorem 4.1 that we have $\theta_2^* = -\infty$, which means that w.p.1 a large queue does not built up in the second queue.

Finally, we found that the departure process D_2 from the second queue has large deviations rate function $\Lambda^{*-}_{D_2}(a)$ equal to $\Lambda^{*-}_{D}(a)$. This, can also be seen analytically. Namely, observe that in Eq. (77) we have $\Lambda^{*-}_{\Gamma}(a) = \Lambda^{*-}_{D}(a)$ which implies $\Lambda^{*-}_{D_2}(a) = \Lambda^{*-}_{D}(a)$.

9 Conclusions and Open Problems

We considered a single class, acyclic network of G/G/1 queues, and characterized the large deviations behaviour of the waiting time and the queue length in all the queues of the network. We accomplished that by obtaining the large deviations behaviour of all the processes resulting from various operations in the network, which for the network model that we considered were passage-from-a-queue, superposition of independent processes, and Bernoulli splitting of a process to a number of processes. We concretely characterized the way that these large deviations occur.

These results are to the best of our knowledge among the few that study large deviations in a network. It is clear that more work is needed in this area, especially in view of the important applications in high speed communication networks. It is an interesting open problem to derive similar results for network models that have feedback and accommodate more than one types of traffic. It would also be interesting to study, in the large deviations regime, how different types of traffic interact and how to choose scheduling policies in order to satisfy certain performance criteria. Work relevant to the latter problem, for the single queue case is reported in [Tse94] and [dVK94].

Acknowledgments

The authors wish to thank G. de Veciana, C. Courcoubetis, J. Walrand and T. Zajic for making available preprints of their work and D. Tse and O. Zeitouni for some helpful discussions.

References

[Ana88] V. Anantharam, *How large delays build up in a GI/G/1 queue*, Queueing Systems **5** (1988), 345–368.

[Asm82] S. Asmussen, *Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the GI/G/1 queue*, Advances in Applied Probability **14** (1982), 143–170.

[BM92] D. Bertsimas and G. Mourtzinou, *A unified method to analyze overtake free queueing systems*, Working paper, Operations Research Center, MIT, 1992.

[BN90] D. Bertsimas and D. Nakazato, *The departure process from a GI/G/1 queue and its applications to the analysis of tandem queues*, Working paper OR 245-91, Operations Research Center, MIT, 1990.

[BN91] D. Bertsimas and D. Nakazato, *The general distributional Little's law and its applications*, Working paper, Operations Research Center, MIT, 1991, To appear, *Operations Research.*

[Buc90] J. A. Bucklew, *Large deviation techniques in decision, simulation, and estimation*, Wiley, New York, 1990.

[Cha94] C.S. Chang, *Stability, queue length and delay of deterministic and stochastic queueing networks*, IEEE Transactions on Automatic Control **39** (1994), no. 5, 913–931.

[Cru91a] R. L. Cruz, *A calculus for network delay, Part I: Network elements in isolation*, IEEE Transactions on Information Theory **37** (1991), no. 1, 114–131.

[Cru91b] R. L. Cruz, *A calculus for network delay, Part II: Network analysis*, IEEE Transactions on Information Theory **37** (1991), no. 1, 132–141.

[CW93] C. Courcoubetis and R. Weber, *Effective bandwidths for stationary sources*, Preprint, 1993.

[dVCW93] G. de Veciana, C. Courcoubetis, and J. Walrand, *Decoupling bandwidths for networks: A decomposition approach to resource management*, Memorandum, Electronics Research Laboratory, U.C. Berkeley, 1993.

[dVK94] G. de Veciana and G. Kesidis, *Bandwidth allocation for multiple qualities of service using Generalized Processor Sharing*, Technical report scc-94-01, Systems Communications & Control group, Department of Electrical and Computer Engineering, The University of Texas at Austin, 1994.

[dVW92] G. de Veciana and J. Walrand, *Traffic shaping for ATM networks: Asymptotic analysis and simulations*, Preprint, 1992.

[dVW93] G. de Veciana and J. Walrand, *Effective bandwidths: Call admission, traffic policing & filtering for ATM networks*, Memorandum, Electronics Research Laboratory, U.C. Berkeley, 1993.

[DZ93a] A. Dembo and T. Zajic, *Large deviations: From empirical mean and measure to partial sums processes*, Preprint, 1993.

[DZ93b] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, Jones and Bartlett, 1993.

[EM93] A. I. Elwalid and D. Mitra, *Effective bandwidth of general Markovian traffic sources and admission control of high speed networks*, IEEE/ACM Transactions on Networking **1** (1993), no. 3, 329–343.

[GA94] A. Ganesh and V. Anantharam, *The stationary tail probability of an exponential server tandem fed by renewal arrivals*, Talk given at the IMA workshop, Minesota, 1994.

[GH91] R.J. Gibbens and P.J. Hunt, *Effective bandwidths for the multi-type UAS channel*, Queueing Systems **9** (1991), 17–28.

[Hui88] J. Y. Hui, *Resource allocation for broadband networks*, IEEE Journal on Selected Areas in Communications **6** (1988), no. 9, 1598–1608.

[Kel79] F.P. Kelly, *Reversibility and stochastic networks*, Wiley, New York, 1979.

[Kel91] F. P. Kelly, *Effective bandwidths at multi-class queues*, Queueing Systems **9** (1991), 5–16.

[Kin70] J.F.C. Kingman, *Inequalities in the theory of queues*, Journal of the Royal Statistical Society **32** (1970), 102–110.

[KWC93] G. Kesidis, J. Walrand, and C.S. Chang, *Effective bandwidths for multiclass Markov fluids and other ATM sources*, IEEE/ACM Transactions on Networking **1** (1993), no. 4, 424–428.

[Pas] I. Ch. Paschalidis, *The large deviations behaviour of networks of G/G/1 queues with applications in communication networks*, Ph.D. thesis, Massachusetts Institute of Technology, In preparation.

[Roc70] R.T. Rockafellar, *Convex analysis*, Princeton University Press, 1970.

[Tse94] D. Tse, *Variable-rate lossy compression and its effects on communication networks*, Ph.D. thesis, Massachusetts Institute of Technology, 1994.

[Wal88] J. Walrand, *An introduction to queueing networks*, Prentice Hall, 1988.

[YS93] O. Yaron and M. Sidi, *Performance and stability of communication networks via robust exponential bounds*, IEEE/ACM Transactions on Networking **1** (1993), no. 3, 372–385.

Appendix

Here we consider an arbitrary process $\{X_i,\ i \in \mathbb{Z}\}$ that satisfies Assumption B and the following:

For every $\epsilon_1, \epsilon_2, \delta, a > 0$, there exists M_X such that for all $n \geq M_X$

$$e^{-n(\Lambda_X^{*-}(a)+\epsilon_2)} \leq \mathbf{P}[S_{i,j}^X - (j-i+1)a \leq \epsilon_1 n,\ 1 \leq i \leq j \leq n \text{ s.t. } (j-i+1) > \delta n]. \tag{126}$$

Inequality (126) is implied by the results in [DZ93a], under some mild mixing assumptions on the process $\{X_i,\ i \in \mathbb{Z}\}$. We prove that the process $\{X_i,\ i \in \mathbb{Z}\}$ satisfies Assumption C for the service times (Eq. (19)), i.e.,

For every $\epsilon_1, \epsilon_2, a > 0$, there exists M'_X such that for all $n \geq M'_X$

$$e^{-n(\Lambda_X^{*-}(a)+\epsilon_2)} \leq \mathbf{P}[S_{i,j}^X - (j-i+1)a \leq \epsilon_1 n,\ 1 \leq i \leq j \leq n]. \tag{127}$$

Since Assumption C for the arrivals (Eq. (18)) is a weaker version of the above it is also satisfied by the process $\{X_i,\ i \in \mathbb{Z}\}$.

Fix positive ϵ_1, ϵ_2 and a. We have

$$\begin{aligned}
&\mathbf{P}[S_{i,j}^X - (j-i+1)a \leq \epsilon_1 n,\ 1 \leq i \leq j \leq n] = \\
&= \mathbf{P}[S_{i,j}^X - (j-i+1)a \leq \epsilon_1 n,\ 1 \leq i \leq j \leq n \text{ s.t. } (j-i+1) > \delta n, \\
&\qquad S_{i,j}^X - (j-i+1)a \leq \epsilon_1 n,\ 1 \leq i \leq j \leq n \text{ s.t. } (j-i+1) \leq \delta n] \\
&\geq \mathbf{P}[S_{i,j}^X - (j-i+1)a \leq \epsilon_1 n,\ 1 \leq i \leq j \leq n \text{ s.t. } (j-i+1) > \delta n] - \\
&\qquad \mathbf{P}[\exists\ i \leq j \in [1,n] \text{ s.t. } (j-i+1) \leq \delta n \text{ and } S_{i,j}^X - (j-i+1)a \geq \epsilon_1 n].
\end{aligned} \tag{128}$$

where we have used the inequality $\mathbf{P}[A \cap B] \geq \mathbf{P}[A] - \mathbf{P}[B^C]$. Using the union bound and the Gärtner-Ellis Thm. we obtain that for all $\epsilon_3 > 0$ there exists N_1 such that for all $n \geq N_1$

$$\begin{aligned}
&\mathbf{P}[\exists\ i \leq j \in [1,n] \text{ s.t. } (j-i+1) \leq \delta n \text{ and } S_{i,j}^X - (j-i+1)a \geq \epsilon_1 n] \leq \\
&\qquad \leq \sum_{\substack{i \leq j \in [1,n] \\ (j-i+1) \leq \delta n}} \mathbf{P}[S_{i,j}^X - (j-i+1)a \geq \epsilon_1 n] \\
&\qquad \leq \sum_{\substack{i \leq j \in [1,n] \\ (j-i+1) \leq \delta n}} \mathbf{P}[S_{1,\delta n}^X \geq \epsilon_1 n] \\
&\qquad \leq n^2 e^{-n\delta(\Lambda_X^{*+}(\frac{\epsilon_1}{\delta})-\epsilon_3)}.
\end{aligned} \tag{129}$$

Now for given $\epsilon'_2 > 0$ choose ϵ_3 and δ small enough in order for large n to have

$$n^2 e^{-n\delta(\Lambda_X^{*+}(\frac{\epsilon_1}{\delta})-\epsilon_3)} \leq \frac{1}{2} e^{-n(\Lambda_X^{*-}(a)+\epsilon'_2)} \tag{130}$$

This can be done since $\Lambda_X^{*+}(\beta) \to \infty$ as $\beta \to \infty$.

Also, by using (126) we have that there exists N'' such that for all $n \geq N''$

$$\mathbf{P}[S_{i,j}^X - (j-i+1)a \leq \epsilon_1 n,\ 1 \leq i \leq j \leq n \text{ s.t. } (j-i+1) > \delta n] \geq e^{-n(\Lambda_X^{*-}(a)+\epsilon_2')} \quad (131)$$

Combining (131), (130) and (129) with (128) we obtain that there exists $\hat{N}$ such that for all $n \geq \hat{N}$

$$\mathbf{P}[S_{i,j}^X - (j-i+1)a \leq \epsilon_1 n,\ 1 \leq i \leq j \leq n] \geq \frac{1}{2} e^{-n(\Lambda_X^{*-}(a)+\epsilon_2')} \quad (132)$$

Finally, to obtain (127) it suffices to choose ϵ_2' such that for large enough n

$$\frac{1}{2} e^{-n(\Lambda_X^{*-}(a)+\epsilon_2')} \geq e^{-n(\Lambda_X^{*-}(a)+\epsilon_2)}$$

■